J. Mol. Biol. (1983) **166**, 477–535

Complete Nucleotide Sequence of Bacteriophage T7 DNA and the Locations of T7 Genetic Elements

JOHN J. DUNN AND F. WILLIAM STUDIER

Biology Department, Brookhaven National Laboratory Upton, N.Y. 11973, U.S.A.

(Received 11 August 1982)

The complete nucleotide sequence of bacteriophage T7 DNA, 39.936 base-pairs, has been determined by the techniques of Maxam & Gilbert. All previously known T7 genes and several unsuspected genes have been identified in the sequence. T7 DNA carries genetic information very efficiently: the coding sequences of 50 genes are close-packed but essentially not overlapping, and occupy almost 92% of the nucleotide sequence. This arrangement strongly suggests that all 50 of these close-packed genes are expressed, although there is as yet evidence for expression of only 38 of them. In addition, five potential overlapping genes have been identified, and there is preliminary evidence that one of them is expressed. Where gaps between coding sequences are found, they usually are less than 100 base-pairs long, and usually contain one or more transcription signals, RNAase III cleavage sites, or origins of replication. Transcription signals in the T7 DNA include the three strong early promoters and the early termination site for *Escherichia coli* RNA polymerase, and 17 promoters and one termination site for T7 RNA polymerase. Ten RNAase III cleavage sites have been located, five in the early region and five in the late region. The primary transcripts are processed at these sites to provide the messenger RNAs observed *in vivo*. Almost all of the T7 messenger RNAs are polycistronic, but there are few polar effects at the level of transcription or translation, and most T7 proteins seem to be initiated independently, each from its own ribosome-binding and initiation site. The initiation codon for most T7 proteins is AUG, but a few proteins are predicted to begin at GUG. Certain T7 genes specify pairs of overlapping proteins. The two proteins specified by gene *4* are made in about equal amounts, beginning at two different ribosome-binding and initiation sites in the same reading frame and ending at a common termination codon. The two proteins specified by gene *10* are made in very different amounts. They begin at the same initiation site, but the minor gene *10* protein appears to be produced by a shift in translational reading frame just ahead of the normal termination codon, thereby adding 53 amino acids to the COOH-terminal end of the major protein. Gene *10* specifies the major capsid protein of the phage particle, and both the major and minor gene *10* proteins are incorporated into the phage particle. One or two other T7 genes appear to utilize translational frameshifting to produce unequal amounts of proteins that differ at their COOH-terminal ends. The amino acid sequences and compositions predicted for all of the T7 proteins (except the proteins produced by frameshifting) are given. T7 DNA begins and ends with a perfect direct repeat of 160 base-pairs. Immediately adjacent to this terminal repetition, at both ends of the mature DNA, lie very similar, regular arrays of 12 imperfect copies of a seven-

0022-2836/83/160477-59 $03.00/0

base sequence. These arrays occupy about 160 base-pairs, starting about 15 base-pairs from the terminal repetition. In the concatemeric form of T7 DNA, a single copy of the terminal repetition is flanked by these two arrays of repeated sequences, and it seems likely that this arrangement is involved somehow in formation of the ends of mature T7 DNA.

1. Introduction

Bacteriophage T7 and its DNA have been the objects of considerable genetic and biochemical investigation, and the nucleotide sequence of the first 30% of T7 DNA has been reported (Dunn & Studier, 1981; Stahl & Zinn, 1981). We have now completed the determination of the sequence of the entire 39,936 base-pairs of T7 DNA, and have identified the positions of individual genes and genetic signals in the sequence.

2. Determination of the Nucleotide Sequence

The nucleotide sequence was determined on DNA from purified phage particles of wild-type T7 (Studier, 1969,1979) by the methods of Maxam & Gilbert (1977,1979), using previously determined restriction maps of T7 DNA as the starting point for isolating specific DNA fragments (McDonell *et al.*, 1977; Rosenberg *et al.*, 1979; unpublished results). Computer programs for storing, searching and analyzing the nucleotide sequence were developed and applied by K. Thompson and W. Crockett.

It is often convenient to give position in the T7 DNA molecule in units of 1% the total length of T7 DNA, beginning at the genetic left end. Previously, we used a value of 400 bases or base-pairs for a T7 unit, because we estimated that T7 DNA would be close to 40,000 base-pairs long. The nucleotide sequence reported here gives a value of 399·36 for a T7 unit, which is the value used throughout this paper.

The sequence given in Fig. 1 is that of the *l*-strand of T7 DNA, the strand that has its 5′ phosphate at the left end of the genetic map and has the same sequence as the T7 messenger RNAs. In the sections that follow, the transcription and translation signals in T7 DNA will be discussed in detail. Tables 1 to 3 give the locations of these signals in the nucleotide sequence, and can be used in conjunction with Fig. 1 to locate the sequences referred to.

The sequence of nucleotides 12,101 to 39,936 is newly reported in this paper. In this region, the sequence of 24,166 nucleotides was determined in the *l* strand and of 23,962 nucleotides in the *r* strand, which means that 72·9% of the sequence was determined in both strands. The sequences across all of the restriction sites used to end-label fragments for sequencing were overlapped from another site.

All of the nucleotide sequence of Fig. 1 except for nucleotides 3282 to 5902 was determined by one of us (J.J.D.), who proof-read the entire set of sequencing films against printouts of the computer file of the sequence. After all corrections had been made, the computer file was used to print Fig. 1. There is nothing that we currently know about the genetics of T7 or the physical properties of T7 DNA that is in disagreement with the nucleotide sequence given in Fig. 1, and we expect that the error frequency in this sequence is very low.

The sequence for nucleotides 3283 to 5901 is from Stahl & Zinn (1981) and Oakley & Coleman (1977). Grachev & Pletnev (1981) have also reported a sequence for nucleotides 2858 to 5855, but their sequence differs from that of Fig. 1 in 80 places. We use the sequences of Stahl & Zinn and of Oakley & Coleman in the regions we have not determined because, in the regions that overlap, there are several differences between our sequence and that of Grachev & Pletnev but none between our sequence and that of Stahl & Zinn. Furthermore, restriction patterns predicted by the sequence of Grachev & Pletnev for some

```
    0  TCTCACACTG TACGGACCTA AAGTTCCCCC ATAGGGGGTA CCTAAAGCCC AGCCAATCAC CTAAAGTCAA CCTTCGGTTG ACCTTCAGGG TTCCCTAAGG
       GTTGGGGATG ACCCTTGGGT TTGTCTTTGG GTGTTACCTT GAGTGTCTCT CTGTGTCCCT ATCTGTTACA GTCTCCTAAA GTATCCTCCT AAAGTCACCT
       CCTAACGTCC ATCCTAAAGC CAACACCTAA AGCCTACACC TAAAGACCCA TCAAGTCAAC GCCTATCTTA AAGTTTAAAC ATAAAGACCA GACCTAAAGA
       CCAGACCTAA AGACACTACA TAAAGACCAG ACCTAAAGAC GCCTTGTTGT TAGCCATAAA GTGATAACCT TTAATCATTG TCTTTATTAA TACAACTCAC
       TATAAGGAGA GACAACTTAA AGACACTTAA AAGATTAATT TAAAATTTAT CAAAAAGAGT ATTGACTTAA AGTCTAACCT ATAGGATACT TACAGCCATC
  500  GAGAGGGACA CGGCGAATAG CCATCCCAAT CGACACCGGG GTCAACCGGA TAGTAGACA GCCTGATAAG TCGCACGAAA AACAGGTATT GACAACATGA
       AGTAACATGC AGTAAGATAC AAATCGCTAG GTAACACTAG CAGCGTCAAC CGGGCGCACA GTGCCTTCTA GGTGACTTAA GCGCACCACG GCACATAAGG
       TGAAACAAAA CGGTTGACAA CATGAAGTAA ACACGGTACG ATGTACCACA TGAAACGACA GTGAGTCACC ACACTGAAAG GTGATGCGGT CTAACGAAAC
       CTGACCTAAG ACGTCTTTTA ACAATCTGGT AAATAGCTCT TGAGTGCATG ACTAGCGGAT AACTCAAGGG TATCGCAAGG TGCCCTTTAT GATATTCACT
       AATAACTGCA CGAGGTAACA CAACATGGCT ATGTCTAACA TGACTTACAA CAACGTTTTC GACCACGCTT ACGAAATGCT GAAACAAAAC ATCCGTTATG
  1000 ATGACATCCG TGACACTGAT GACCTGCACG ATGCTATTCA CATGGCTGCC GATAATGCAG TTCCGCACTA CTACGCTGAC ATCTTTAGCG TAATGGCAAG
       TCAGGGCATT GACCTTGAGT TCGAAGACTC TGGTCTGATG CCTGACACCA AGGAGTAAAT CCGCATCCTG CAAGCGCGTA TCTATGAGCA ATTAACGATT
       GACCTTGGCG AAGACGGAGA AGACTTGCTT AATGAATACT TGGAGGAAGT CGAGGAGTAC GAGGAGGATG AAGAGTAATG TCTACTACCA ACGTGCAATA
       CGGTCTGACC GCTCAAATTG TACTTTCTTA TACGCACATG GTGCGCTGTG GCTTAACTG GTCACTCGCA ATGGCACAGC TCAAAGAACT GTACGAAAAC
       AACAAGGCAA TACCTTTAGA ATCTGCTGAG TGATAGACTC AAGGTCGCTC CTACCGAGTG GCCTTTATGA TTATCACTTT ACTTATGAGG GAGTAATGTA
  1500 TATGCTTACT ATCGGTCTAC TCACCGCTCT AGGTCTAGCT GTAGGTGCAT CCTTTGGGAA GGCTTTAGGT GTAGCTGTAG GTTCCTACTT TACCGCTTGC
       ATCATCATAG GAATCATCAA AGGGGCACTA CGCAAATGAT GAAGCACTAC GTTATGCCAA TCCACACGTC CAACGGGGCA ACCGTATGTA CACCTGATGG
       GTTCGCAATG AAACAACGAA TCGAACGCCT TAAGCGTGAA ETCCGCATTA ACCGCAAGAT TAACAAGATA GGTTCCGGCT ATGACAGAAC GCACTGATGC
       CTTAAAGAAA GGTTATATGC CCAATGGCAC ACTATACGCT GCAAATCGGC GAATAGTGAC AACTTGGCGA GAGAACAACC TCGAACGCCG CAAGGACAAG
       AGAGGGCGGC GTGGCATAGA CGAAAGGAAA AGGTTAAAGC CAAGAAACTC GCCGCACTTG AACAGGCACT AGCCAACACA CTGAACGCTA TCTCATAACG
  2000 AACATAAAGG ACACAATGCA ATGAACATTA CCGACATCAT GAACGCTATC AAGCACTGCC AATCTGTGAA CTTGACAAGC GTCAAGGTAT
       GCTTATCGAC TTACTGGTCG AGATGGTCAA CAGCGAGACG TGTCATGGCG AGCTAACCGA ACTAAATCAG GCACTTGAGC ATCAAGATTG GTGGACTACC
       TTGAAGTGTC TCACGGCTGA CGCAGGGTTC AAGATGCTCG GTAATGGTCA CTTCTCGGCT GCTTATAGTC ACCCGCTGCT ACCTAACAGA GTGATTAAGG
       TGGGCTTTAA GAAAGAGGAT TCAGGCGCAG CCTATACCGC ATTCTGCCGC ATGTATCAGG GTCGTCCTGG TATCCCTAAC GTCTACGATG TACAGCGCCA
       CGCTGGATGC TATACGGTGG TACTTGACGC ACTTAAGGAT TGCGAGCGTT TCAACAATGA TGCCCATTAT AAATACGCTG AGATTGCAAG CGACATCATT
  2500 GATTGCAATT CGGATGAGCA TGATGAGTTA ACTGGATGGG ATGGTGAGTT TGTTGAAACT TGTAAACTAA TCCGCAAGTT CTTTGAGGGC ATCGCCTCAT
       TCGACATGGA TAGCGGGAAC ATCATGTTCT CAAATGGAGA CGTACCATAC ATCACCGACC CGGTATCATT CTCGCAGAAG AAAGACGGTG GCGCATTCAG
       CATCGACCCT GAGGAACTCA TCAAGGAAGT CGAGGAAGTC GCACGACAGA AAGAAATTGA CCGCGCTAAG GCCCGTAAAG AACGTCACGA GGGCGCTTA
       GAGGCACGCA GATTCAAACG TCGCAACCGC AAGGCACGTA AAGCACACAA AGCTAAGCGC GAAAGAATGC TTGCTGCGTG GCGATGGGCT GAACGTCAAG
       AACGGCGTAA CCATGAGGTA GCTGTAGATG TACTAGGAAG AACCAATAAC GCTATGCTCT GGGTCAACAT GTTCTCTGGG GACTTTAAGG CGCTTGAGGA
  3000 ACGAATCGCG CTGCACTGGC GTAATGCTGA CCGGATGGCT ATCGCTAATG GTCTTACGCT CAACATTGAT AAGCAACTTG ACGCAATGTT AATGGGCTGA
       TAGTCTTATC TTACAGGTCA TCTGCGGGTG GCCTGAATAG GTACGATTTA CTAACTGGAA GAGGCACTAA ATGAACACGA TTAACATCGC TAACGAACGAC
       TTCTCTGACA TCGAACTGGC TGCTATCCCG TTCAACACTC TGGCTGACCA TTACGGTGAG CGTTTAGCTC GCGAACAGTT GGCCCTTGAG CATGAGTCTT
       ACGAGATGGG TGAAGCACGC TTCCGCAAGA TGTTTGAGCG TCAACTTAAA GCTGGTGAGG TTGCGGATAA CGCTGCCGCC AAGCCTCTCA TCACTACCCT
       ACTCCCTAAG ATGATTGCAC GCATCAACGA CTGGTTTGAG GAAGTGAAAG CTAAGCGCGG CAAGCGCCCG ACAGCCTTCC AGTTCCTGCA AGAAATCAAG
  3500 CCGGAAGCCG TAGCGTACAT CACCATTAAG ACCACTCTGG CTTGCCTAAC CAGTGCTGAC AATACAACCG TTCAGGCTGT AGCAAGCGCA ATCGGTCGGG
       CCATTGAGGA CGAGGCTCGC TTCGGTCGTA TCCGTGACCT TGAAGCTAAG CACTTCAAGA AAAACGTTGA GGAACAACTC AACAAGCGCG TAGGGCACGT
       CTACAAGAAA GCATTTATGC AAGTTGTCGA GGCTGACATG CTCTCTAAGG GTCTACTCGG TGGCGAGGCG TGGTCTTCGT GGCATAAGGA AGACTCTATT
       CATGTAGGAG TACGCTGCAT CGAGATGCTC ATTGAGTCAA CCGGAATGGT TAGCTTACAC CGCCAAAATC CTGGCGTAGT AGGTCAAGAC TCTGAGACTA
       TCGAACTCGC ACCTGAATAC GCTGAGGCTA TCGCAACCCG TGCAGGTGCG CTGGCTGGCA TCTCTCCGAT GTTCCAACCT TGCGTAGTTC CTCCTAAGCC
  4000 GTGGACTGGC ATTACTGGTG GTGGCTATTG GGCTAACGGT CGTCGTCCTC TGGCGCTGGT GCGTACTCAC AGTAACAAAG CACTGATGCG CTACGAAGAC
       GTTTACATGC CTGAGGTGTA CAAAGCGATT AACATTGCGC AAAACACCGC ATGGAAAATC AACAAGAAAG TCCTAGCGGT CGCCAACGTA ATCACCAAGT
       GGAAGCATTG TCCGGTCGAG GACATCCCTG CGATTGAGCG TGAAGAACTC CCGATCGAAAC CGGAAGACAT CGACATGAAT CCTGAGGCTC TCACCGCGTG
       GAAACGTGCT GCCGCTGCTG TGTACCGCAA GACAAGGCTC GCAAGTCTCG CCGTATCAGC CTTGAGTTCA TGCTTGAGCA AGCCAATAAG TTTGCTAACC
       ATAAGGCCAT CTGGTTCCCT TACAACATGG ACTGGCGCGG TTCGTGTTTA CGCTGTGTCA ATGTTCAACC CGCAAGGTAA CGATATGACC AAAGGACGTC
  4500 TTACGCTGGC GAAAGGTAAA CCAATCGGTA AGGAAGGTTA CTACTGGCTG AAAATCCACG GTGCAAACTG TGCGGGTGTC GATAAGGTTT CGTTTCCTGA
       GCGCATCAAG TTCATTGAGG AAAACCACGA GAACATCATG GCTTGCGCTA AGTCTCCACT GGAGAACACT TGGTGGGCTG AGCAAGATTC TCCGTTCTGC
       TTCCTTGCGT TCTGCTTTGA GTACGCTGGG GTACAGCACC ACGGCCTGAG CTATAACTGC TCCCTTCCGC TGGCGTTTGA CGGGTCTTGC TCTGGCATCC
       AGCACTTCTC CGCGATGCTC CGACATGAGG TAGGTGGTCG CGCGGTTAAC TTGCTTCCTA GTGAAACCGT TCACGACATC TACGGGATTG TTGCTAACAA
       AGTCAACGAG ATTCTGCAAG CACACGCAAT CAATGGGACC GATAACGAAG TAGTTACCGT GACCGATGAG AACACTGGTG AAATCTCTGA GAAAGTCAAG
```

FIG. 1(a)

```
5000  CTGGGCACTA AGGCACTGGC TGGTCAATGG CTGGCTTACG GTGTTACTCG CAGTGTGACT AAGCGTTCAG TCATGACGCT GGCTTACGGG TCCAAAGAGT
      TCGGCTTCCG TCAACAAGTG CTGGAAGATA CCATTCAGCC AGCTATTGAT TCCGGCAAGG GTCTGATGTT CACTCAGCCG AATCAGGCTG CTGGATACAT
      GGCTAAGCTG ATTTGGGAAT CCGTCAGCGT GACGGTGGTA GCTGCGGTTG AAGCAATGAA CTGGCTTAAG TCTGCTGCTA AGCTGCTGGC TGCTGAGGTC
      AAAGATAAGA AGACTGGAGA GATTCTTCGC AAGCGTTGCG CTGTGCATTG GGTAACTCCT GATGGTTTCC CTGTGTGGCA GCAATACAAG AAGCCTATTC
      AGACGCGCTT GAACCTGATG TTCCTCGGTC AGTTCCGCTT ACAGCCTACC ATTAACACCA ACAAAGATAG CGAGATTGAT GCACACAAAC AGGAGTCTGG
5500  TATCGCTCCT AACTTTGTAC ACAGCCAAGA CGGTAGCCAC CTTCGTAAGA CTGTAGTGTG GGCACACGAG AAGTACGGAA TCGAATCTTT TGCACTGATT
      CACGACTCCT TCGGTACCAT TCCGGCTGAC GCTGCGAACC TGTTCAAAGC AGTGCGCGAA ACTATGGTTG ACACATATGA GTCTTGTGAT GTACTGGCTG
      ATTTCTACGA CCAGTTCGCT GACCAGTTGC ACGAGTCTCA ATTGGACAAA ATGCCAGCAC TTCCGGCTAA AGGTAACTTG AACCTCCGTG ACATCTTAGA
      GTCGGACTTC GCGTTCGCGT AACGCCAAAT CAATACGACT CACTATAGAG GGACAAACTC AAGGTCATTC GCAAGAGTGG CCTTTATGAT TGACCTTCTT
      CCGGTTAATA CGACTCACTA TAGGAGAACC TTAAGGTTTA ACTTTAAGAC CCTTAAGTGT TAATTAGAGA TTTAAATTAA AGAATTACTA AGAGAGGACT
6000  TTAAGTATGC GTAACTTCGA AAAGATGACC AAACGTTCTA ACCGTAATGC TCGTGACTTC GAGGCAACCA AAGGTCGCAA GTTGAATAAG ACTAAGCGTG
      ACCGCTCTCA CAACCGTAGC TGGGAGGGTC AGTAAGATGG GACGTTTATA TAGTGGTAAT CTGGCAGCAT TCAAGGCAGC AACAAACAAG CTGTTCCAGT
      TACACTTAGC GGTCATTTAT GATGACTGGT ATGATGCCTA TACAAGAAAA GATTGCATAC GGTTACGTAT TCAGGACAGG AGTGGAAACC TGATTGATAC
      TACCACCTTC TACCACCACG ACGAGGACGT TCTGTTCAAT ATGTGTACTG ATTGGTTGAA CCATATGTAT GACCAGTTGA AGGACTGGAA GTAATACGAC
      TCAGTATAGG GACAATGCTT AAGGTCGCTC TCTAGGAGTG GCCTTAGTCA TTTAACCAAT AGGAGATAAA CATTATGATG AACATTAAGA CTAACCCGTT
6500  TAAAGCCGTG TCTTTCGTAG AGTCTGCCAT TAAGAAGGCT CTGGATAACG CTGGGTATCT TATCGCTGAA ATCAAGTACG ATGGTGTACG CGGCAACATC
      TGCGTAGACA ATACTGCTAA CAGTTACTGG CTCTCTCGTG TATCTAAAAC GATTCCGGCA CTGGAGCACT TAAACGGGTT TGATGTTCGC TGGAAGCGTC
      TACTGAACGA TGACCGTTGC TTCTACAAAG ATGGCTTTAT GCTTGATGGG GAACTCATGG TCAAGGGCGT AGACTTTAAC ACAGGGTCCG GCCTACTGCG
      TACCAAATGG ACTGACACGA AGAACCAAGA GTTCCATGAA GAGTTATTCG TTGAACCAAT CCGTAAGAAA GATAAAGTTC CCTTTAAGCT GCACACTGGA
      CACCTTCACA TAAAACTGTA CGCTATCCTC CCGCTGCACA TCGTGGAGTC TGGAGAAGAC TGTGATGTCA TGACGTTGCT CATGCAGGAA CACGTTAAGA
7000  ACATGCTGCC TCTGCTACAG GAATACTTCC CTGAAATCGA ATGGCAAGCG GCTGAATCTT ACGAGGTCTA CGATATGGTA GAACTACAGC AACTGTACGA
      GCAGAAGCGA GCAGAAGGCC ATGAGGGTCT CATTGTGAAA GACCCGATGT GTATCTATAA GCGCGGTAAG AAATCTGGCT GGTGGAAAAT GAAACCTGAG
      AACGAAGCTG ACGGTATCAT TCAGGGTCTG GTATGGGGTA CAAAAGGTCT GGCTAATGAA GGTAAAGTGA TTGGTTTTGA GGTGCTTCTT GAGAGTGGTC
      GTTTAGTTAA CGCCACGAAT ATCTCTCGCG CCTTAATGGA TGAGTTCACT GAGACAGTAA AAGAGGCCAC CCTAAGTCAA TGGGCATTCT TTAGCCCATA
      CGGTATTGGC GACAACGATG CTTGTACTAT TAACCCTTAC GATGGCTGGG CGTGTCAAAT TAGCTACATG GAGGAAACAC CTGATGGCTC TTTGCGGCAC
7500  CCATCGTTCG TAATGTTCCG TGGCACCGAG GACAACCCTC AAGAGAAAAT GTAATCACAC TGGCTCACCT TCGGGTGGGC CTTTCTGCGT TTATAGGAG
      ACACTTTATG TTTAAGAAGG TTGGTAAATT CCTTGCGGCT TTGGCAGCTA TCCTGACGCT TGCGTATATT CTTGCGGTAT ACCCTCAAGT AGCACTAGTA
      GTAGTTGGCG CTTGTTACTT AGCGGCAGTG TGTGCTTGCG TGTGGAGTAT AGTTAACTGG TAATACGACT CACTAAAGGA GGTACACACC ATGATGTACT
      TAATGCCATT ACTCATCGTC ATTGTAGGAT GCCTTGCGCT CCACTGTAGC GATGATGATA TGCCAGATGG TCACGCTTAA TACGACTCAC TAAAGGAGAC
      ACTATATGTT TCGACTTCAT TACAACAAAA GCGTTAAGAA TTTCACGGTT CGCCGTGCTG ACCGTTCAAT CGTATGTGCG AGCGAGCGCC GAGCTAAGAT
8000  ACCTCTTATT GGTAACACAG TTCCTTTGGC ACCGAGCGTC CACATCATTA TCACCCGTGG TGACTTTGAG AAAGCAATAG ACAAGAAACG TCCGGTTCTT
      AGTGTGGCAG TGACCCGCTT CCCGTTCGTC CGTCTGTTAC TCAAACGAAT CAAGGAGGTG TTCTGATGGG ACTGTTAGAT GGTGAAGCCT GGGAAAAAGA
      AAACCCGCCA GTACAAGCAA CTGGGTGTAT AGCTTGCTTA GAGAAAGATG ACCGTTATCC ACACACCTGT AACAAAGGAG CTAACGATAT GACCGAACGT
      GAACAAGAGA TGATCATTAA GTTGATAGAC AATAATGAAG GTCGCCCAGA TGATTTGAAT GGCTGCGGTA TTCTCTGCTC CAATGTCCCT TGCCACCTCT
      GCCCCGCAAA TAACGATCAA AAGATAACCT TAGGTGAAAT CCGAGCGATG GACCCACGTA AACCACATCT GAATAAACCT GAGGTAACTC CTACAGATGA
8500  CCAGCCTTCC GCTGAGACAA TCGAAGGTGT CACTAAGCCT TCCCACTACA TGCTGTTTGA CGACATTGAG GCTATCGAAG TGATTGCTCG TTCAATGACC
      GTTGAGCAGT TCAAGGCATA CTGCTTCGGT AACATCTTAA AGTACAGACT ACGTGCTGGT AAGAAGTCAG AGTTAGCGTA CTTAGAGAAA GACCTAGCGA
      AAGCAGACTT CTATAAAGAA CTCTTTGAGA AACATAAGGA TAAATGTTAT GCATAACTTC AAGTCAACCC CACCTGCCGA CAGCCTATCT GATGACTTCA
      CATCTTGCTC AGAGTGGTGC CGAAAGATGT GGGAAGAGAC ATTCGACGAT GCGTACATCA AGCTGTATGA ACTTTGGAAA TCGAGAGGTC AATGACTATG
      TCAAACGTAA ATACAGGTTC ACTTAGTGTG GACAATAAGA AGTTTTGGGC TACCGTAGAG TCCTCGGAGC ATTCCTTCGA GGTTCCAATC TACGCTGAGA
9000  CCCTAGACGA AGCTCTGGAG TTAGCCGAAT GGCAATACGT TCCGGCTGGC TTTGAGGTTA CTCGTGTGCG TCCTTGTGTA GCACCGAAGT AATACGACTC
      ACTATTAGGG AAGACTCCCT CTGAGAAACC AAACGAAACC TAAAGGAGAT TAACATTATG CCTAAGAAGA TTTTCACCTC TGCGCTGGGT ACCGCTGAAC
      CTTACGCTTA CATCGCCAAG CCGGACTACG GCAACGAAGA GCGTGGCTTT GGGAACCCTC GTGGTGTCTA TAAAGTTGAC CTGACTATTC CCAACAAAGA
      CCCGCGCTGC CAGCGTATGG TCGATGAAAT CGTGAAGTGT CACGAAGAGG CTTATGCTGC TGCCGTTGAG GAATACGAAG CTAATCCACC TGCTGTAGCT
      CGTGGTAAGA ARCCGCTGAA ACCGTATGAG GGTGACATGC CGTTCTTCGA TAACGGTGAC GGTACGACTA CCTTTAAGTT CAAATGCTAC GCGTCTTTCC
9500  AAGACAAGAA GACCAAAGAG ACCAAGCACA TCAATCTGGT TGTGGTTGAC TCAAAAGGTA AGAAGATGGA AGACGTTCCG ATTATCGGTG GTGGCTCTAA
      GCTGAAAGTT AAATATTCTC TGGTTCCATA CAAGTGGAAC ACTGCTGTAG GTGCGACCGT TAAGCTGCAA CTGGAATCCG TGATGCTGCT CGAACTGGCT
      ACCTTTGGTG GCGGTGAAGA CGATTGGGCT GACGAAGTTG AAGAGAACGG CTATGTTGCC TCTGGTTCTG CCAAAGCGAC CAACCACGC GACGAAGAAA
      GCTGGGACGA AGACGACGAA GAGTCCGAGG AAGCAGACGA AGACGGAGAC TTCTAAGTGG AACTGCGGCA GAAAATCCTT GAGCGAATCA AGGTGACTTC
      CTCTGGGTGT TGGGAGTGGC AGGGCGCTAC GAACAATAAA GGGTACGGGC ACGTGTGGTG CAGCAATACC GGAAACGTTG TCTACTGTCA TCGCGTAATG
```

FIG. 1(b)

```
3000  TCTAATGCTC CGAAAGGTTC TACCGTCCTG CACTCCTGTG ATAATCCATT ATGTTGTAAC CCTGAACACC TATCCATAGG AACTCCAAAA GAGAACTCCA
      CTGACATGGT AAATAAGGGT GGCTCACACA AGGGGTATAA ACTTTCAGAC GAAGACGTAA TGGCAATCAT GGAGTCCAGC GAGTCCAATG TATCCTTAGC
      TCGCACCTAT GGTGTCTCCC AACAGACTAT TTGTGATATA CGCAAAGGGA GGCGACATCG CAGGTTACGG CGCTAAAGGA ATCCGAAAGG TTGGAGCGTT
      TCGCTCTGGC CTGAAGGACA AGGTTTCAAA GCAGTTGGAA TCAAAAGGTA TTAAATTCGA GTATGAAGAG TGGAAAGTGC CTTATGTAAT TCCGGCGAGC
      AATCACACTT ACACTCCAGA CTTCTTACTT CCAAACGGTA TATTCGTTGA GACAAAGGGT CTGTGGGAAA GCGATGATAC AAAGAAGCAC TTATTAATTA
3500  GGGAGCAGCA CCCCGAGCTA GACATCCGTA TTGTCTTCTC AAGCTCACGT ACTAAGTTAT ACAAAGGTTC TCCAACGTCT TATGGAGAGT TCTGCGAAAA
      GCATGGTATT AAGTTCGCTG ATAAACTGAT ACCTGCTGAG TGGATAAAGG AACCCAAGAA GGAGGTCCCC TTTGATAGAT TAAAAAGGAA AGGAGGAAAG
      AAATAATGGC TCGTGTACAG TTTAAACAAC GTGAATCTAC TGACGCAATC TTTGTTCACT GCTCGGCTAC CAAGCCAAGT CAGAATGTTG GTGTCCGTGA
      GATTCGCCAG TGGCACAAAG AGCAGGGTTG GCTCGATGTG GGATACCACT TTATCATCAA GCGAGACGGT ACTGTGGAGG CAGGACGAGA TGAGATGGCT
      GTAGGCTCTC ACGCTAAGGG TTACAACCAC AACTCTATCG GCGTCTGCCT TGTTGGTGGT ATCGACGATA AAGGTAAGTT CGACGCTAAC TTTACGCCAG
3000  CCCAAATGCA ATCCCTTCGC TCACTGCTTG TCACACTGCT GGCTAAGTAC GAAGGCGCTG GTCTTCGCGC CCATCATGAG GTGGCGCCGA AGGCTTGCCC
      TTCGTTCGAC CTTAAGCGTT GGTGGGAGAA GAACGAACTG GTCACTTCTG ACCGTGGATA ATTAATTGAA CTCACTAAAG GGAGACCACA GCGGTTTCCC
      TTTGTTCGCA TTGGAGGTCA AATAATGCGC AAGTCTTATA AACAATTCTA TAAGGCTCCG AGGAGGCATA TCCAAGTGTG GGAGGCAGCC AATGGGCCTA
      TACCAAAAGG TTATTATATA GACCACATTG ACGGCAATCC ACTCAACGAC GCCTTAGACA ATCTCCGTCT GGCTCTCCCA AAAGAAAACT CATGGAACAT
      GAAGACTCCA AAGAGCAATA CCTCAGGACT AAAGGGACTG AGTTGGAGCA AGGAAAGGGA GATGTGGAGA GGCACTGTAA CAGCTGAGGG TAAACAGCAT
3500  AACTTTCGTA GTAGAGATCT ATTGGAAGTC GTTGCGTGGA TTTATAGAAC TAGGAGGGAA TTGCATGGAC AATTCGCACG ATTCCGATAG TGTATTTCTT
      TACCACATTC CTTGTGACAA CTGTGGGAGT AGTGATGGGA ACTCGCTGTT CTCTGACGGA CACACGTTCT GCTACGTATG CGAGAAGTGG ACTGCTGGTA
      ATGAAGACAC TAAAGAGAGG GCTTCAAAAC GGAAACCCTC AGGAGGTAAA CCAATGACTT ACAACGTGTG GAACTTCGGG GAATCCAATG GACGCTACTC
      CGCGTTAACT GCGAGAGGAA TCTCCAAGGA AACCTGTCAG AAGGCTGGCT ACTGGATTGC CAAAGTAGAC GGTGTGATGT ACCAAGTGGC TGACTATCGG
      GACCAGAACG GCAACATTGT GAGTCAGAAG GTTCGAGATA AAGATAAGAA CTTTAAGACC ACTGGTAGTC ACAAGAGTGA CGCTCTGTTC GGGAAGCACT
4000  TGTGGAATGG TGGTAAGAAG ATTGTCGTTA CAGAAGGTGA AATCGACATG CTTACCGTGA TGGAACTTCA AGACTGTAAG TATCCTGTAG TGTCGTTGGG
      TCACGGTGCC TCTGCCGCTA AGAAGACATG CGCTGCCAAC TACGAATACT TTGACCAGTT CGAACAGATT ATCTTAATGT TCGATATGGA CGAAGCAGGG
      CGCAAAGCAG TCGAAGAGGC TGCACAGGTT CTACCTGCTG GTAAGGTACG AGTGGCAGTT CTTCCGTGTA AGGATGCAAA CGAGTGTCAC CTAAATGGTC
      ACGACCGTGA AATCATGGAG CAAGTGTGGA ATGCTGGTCC TTGGATTCCT GATGGTGTGG TATCGGCTCT TTCGTTACGT GAACGAATCC GTGAGCACCT
      ATCGTCCGAG GAATCAGTAG GTTTACTTTT CAGTGGCTGC ACTGGTATCA ACGATAAGAC CTTAGGTGCC CGTGGTGGTG AAGTCATTAT GGTCACTTCC
4500  GGTTCCGGTA TGGGTAAGTC AACGTTCGTC CGTCAACAAG CTCTACAATG GGGCACAGCG ATGGGCAAGA AGGTAGGCTT AGCGATGCTT GAGGAGTCCG
      TTGAGGAGAC CGCTGAGGAC CTTATAGGTC TACACAACCG TGTCCGACTG AGACAATCCG ACTCACTAAA GAGAGAGATT ATTGAGAACG GTAAGTTCGA
      CCAATGGTTC GATGAACTGT TCGGCAACGA TACGTTCCAT CTATATGACT CATTCGCCGA GGCTGAGACG GATAGACTGC TCGCTAAGCT GGCCTACATG
      CGCTCAGGCT TGGCCTGTGA CGTAATCATT CTAGACCACA TCTCAATCGT CGTATCCGCT TCTGGTGAAT CCGATGAGCG TAAGATGATT GACAACCTGA
      TGACCAAGCT CAAAGGGTTC GCTAAGTCAA CTGGGGTGGT GCTGGTCGTA ATTTGTCACC TTAAGAACCC AGACAAAGGT AAAGCACATG ACGAAGGTCG
5000  CCCCGTTTCT ATTACTGACC TACGTGGTTC TGCCGCACTA CGCCAACTAT CTGATACTAT TATTGCCCTT GAGCGTAATC AGCAAGGCGA TATGCCTAAC
      CTTGTCCTCG TTCGTATTCT CAAGTGCCGC TTTACTGGTC ATACTGGTAT CGCTGGCTAC ATGGAATACA ACAAGGAAAC CGGATGGCTT GAACCATCAA
      GTTACTCAGG GGAAGAAGAG TCACACTCAG AGTCAACAGA CTGGTCCAAC GACACTGACT TCTGACAGGA TTCTTGATGA CTTTCCAGAC GACTACGAGA
      AGTTTCGCTG GAGAGTCCCA TTCTAATACG ACTCACTAAA GGAGACACAC CATGTTCAAA CTGATTAAGA AGTTAGGCCA ACTGCTGGTT CGTATGTACA
      ACGTGGAAGC CAAGCGACTG AACGATGAGG CTCGTAAAGA GGCCACACAG TCACGCGCTC TGGCGATTCG CTCCAACGAA CTGCCTGACA GTGCATCCAC
5500  TAAAGTTACC GAGGCTGCCC GTGTGGCAAA CCAAGCTCAA CAGCTTTCCA AATTCTTTGA GTAATCAAAC AGGAGAAACC ATTATGTCTA ACGTAGCTGA
      AACTATCCGT CTATCCGATA CAGCTGACCA GTGGAACCGT CGAGTCCACA TCAACGTTCG CAACGGTAAG GCGACTATGG TTTACCGCTG GAAGGACTCT
      AAGTCCTCTA AGAATCACAC TCAGCGTATG ACGTTGACAG ATGAGCAAGC ACTGCGTCTG GTCAATGCGC TTACCAAAGC TGCCGTGACA GCAATTCATG
      AAGCTGGTCG CGTCAATGAA GCTATGGCTA TCCTCGACAA GATTGATAAC TAAGAGTGGT ATCCTCAAGG TCGCCAAAGT GGTGGCCTTC ATGAATACTA
      TTCGACTCAC TATAGGAGAT ATTACCATGC GTGACCCTAA AGTTATCCAA GCAGAAATCG CTAAACTGGA AGCTGAACTG GAGGACGTTA AGTACCATGA
6000  AGCTAAGACT CGCTCCGCTG TTCACATCTT GAAGAACTTA GGCTGGACTT GGACAAGACA GACTGGCTGG AAGAAACCAG AAGTTACCAA GCTGAGTCAT
      AAGGTGTTCG ATAAGGACAC TATGACCCAC ATCAAGGCTG GTGATTGGGT TAAGGTTGAC ATGGGAGTTG TTGGTGGATA CGGCTACGTC CGCTCAGTTA
      GTGGCAAATA TGCACAAGTG TCATACATCA CAGGTGTTAC TCCACGCGGT GCAATCGTTG CCGATAAGAC CAACATGATT CACACAGGTT TCTTGACAGT
      TGTTTCATAT GAAGACATTG TTAGTCACGA ATAATCAATA GGAGAAATCA ATATGATCGT TTCTGACATC GAAGCTAACG CCCTCTTAGA GAGCGTCACT
      AAGTTCCACT GCGGCGTTAT CTACGACTAC TCCACCGCTG AGTACGTAAG CTACCGTCCG AGTGACTTCG GTGCGTATCT GGATGCGCTG GAAGCCGAGG
6500  TTGCACGAGG CGGTCTTATT GTGTTCCACA ACGGTCACAA GTATGACGTT CCTGCATTGA CCAAACTGGC AAAGTTGCAA TTGAACCGAG AGTTCCACCT
      TCCTCGTGAG AACTGTATTG ACACCCTTGT GTTGTCACGT TTGATTCATT CCAACCTCAA GGACACCGAT ATGGGTCTTC TGCGTTCCGG CAAGTTGCCC
      GCAAAACGCT TTGGGTCTCA CGCTTTGGAG GCGTGGGGTT ATCGCTTAGG CGAGATGAAG GGTGAATACA AAGACGACTT TAAGCGTATG CTTGAAGAGC
      AGGGTGAAGA ATACGTTGAC GGAATGGAGT GGTGGAACTT CAACGAAGAG ATGATGGACT ATAACGTTCA GGACGTTGTG GTAACTAAAG CTCTCCTTGA
      CAAGCTACTC TCTGACAAAC ATTACTTCCC TCCTGAGATT GACTTTACGG ACGTAGGATA CACTACGTTC TGGTCAGAAT CCCTTGAGGC CGTTGACATT
```

FIG. 1(c)

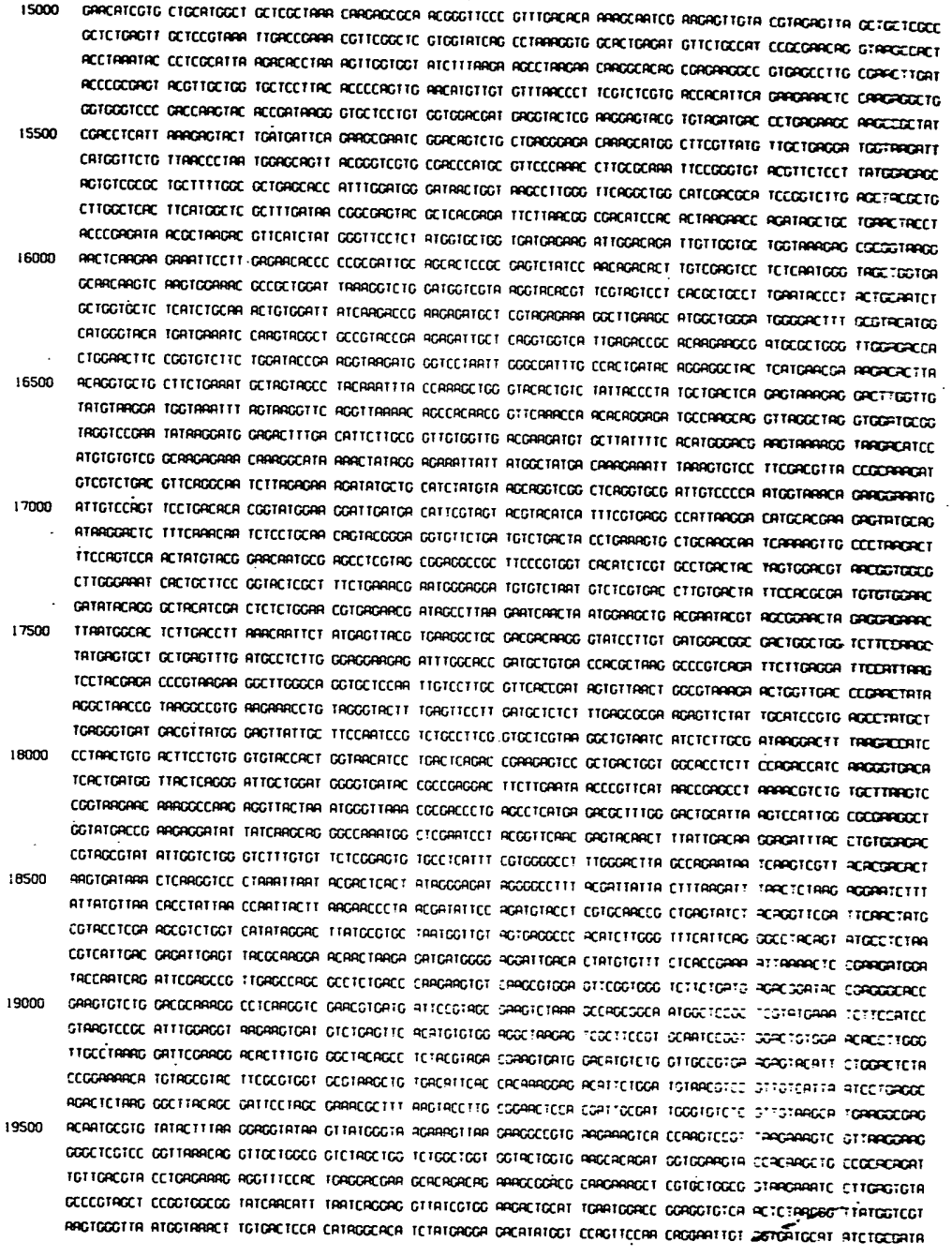


Fig. 1(d)

```
20000  ACCCTAGGTG CTATAACATA AAGCACCTTA CGCTTGGAAC TCCAAAGGAT AATTCCGAGG ACATGGTTAC CAAAGGTAGA CAGGCTAAAG GAGAGGAACT
       AAGCAAGAAA CTTACAGAGT CAGACGTTCT CGCTATACGC TCTTCAACCT TAAGCCACCG CTCCTTAGGA GAACTGTATG GAGTCAGTCA ATCAACCATA
       ACGCGAATAC TACAGCGTAA GACATGGAGA CACATTTAAT GGCTGAGAAA CGAACAGGAC TTGCGGAGGA TGGCGCAAAG TCTGTCTATG AGCGTTTAAA
       GAACGACCGT GCTCCCTATG AGACACGCGC TCAGAATTGC GCTCAATATA CCATCCCATC ATTGTTCCCT AAGGACTCCG ATAACGCCTC TACAGATTAT
       CAAACTCCGT GGCAAGCCGT GGGCGCTCGT GGTCTGAACA ATCTAGCCTC TAAGCTCATG CTGGCTCTAT TCCCTATGCA GACTTGGATG CGACTTACTA
20500  TATCTGAATA TGAAGCAAAG CAGTTACTGA GCGACCCCGA TGGACTCGCT AAGGTCGATG AGGGCCTCTC GATGGTAGAG CGTATCATCA TGAACTACAT
       TGAGTCTAAC AGTTACCGCG TGACTCTCTT TGAGGCTCTC AAACAGTTAG TCGTAGCTGG TAACGTCCTG CTGTACCTAC CGGAACCGGA AGGGTCAAAC
       TATAATCCCA TGAAGCTGTA CCGATTGTCT TCTTATGTGC TCCAACGAGA CGCATTCCGC AACGTTCTGC AAATGGTGAC TCGTGACCAG ATAGCTTTTG
       GTGCTCTCCC TGAGGACATC CGTAAGGCTG TAGAAGGTCA AGGTGGTGAG AAGAAAGCTG ATGAGACAAT CGACGTGTAC ACTCACATCT ATCTGGATGA
       GGACTCAGGT GAATACCTCC GATACGAAGA GGTCGAGGGT ATGGAAGTCC AAGGCTCCGA TGGGACTTAT CCTAAAGAGG CTTGCCCATA CATCCCGATT
21000  CGGATGGTCA GACTAGATGG TGAATCCTAC GGTCGTTCGT ACATTGAGGA ATACTTAGGT GACTTACGGT CCCTTGAAAA TCTCCAAGAG GCTATCGTCA
       AGATGTCCAT GATTAGCTCT AAGGTTATCG GCTTAGTGAA TCCTGCTGGT ATCACCCAGC CACGCCGACT GACCAAAGCT CAGACTGGTG ACTTCGTTAC
       TGGTCGTCCA GAAGAGATCT CGTTCCTCCA ACTGGAGAAG CAAGCAGACT TTACTGTAGC TAAAGCCGTA AGTGACGCTA TCGAGGCTCG CCTTTCGTTT
       GCCTTTATGT TGAACTCTGC GGTTCAGCGT ACAGGTGAAC GTGTGACCGC CGAAGAGATT CGGTATGTAG CTTCTGAACT TGAAGATACT TTAGGTGGTG
       TCTACTCTAT CCTTTCTCAA GAATTACAAT TGCCTCTGGT ACGAGTGCTC TTGAAGCAAC TACAAGCCAC GCAACAGATT CCTGAGTTAC CTAAGGAAGC
21500  CGTAGAGCCA ACCATTAGTA CAGGTCTGGA AGCAATTGGT CGAGGACAAG ACCTTGATAA GCTGGAGCGG TGTGTCACTG CGTGGGCTGC ACTGGCACCT
       ATGCGGGACG ACCCTGATAT TAACCTTGCG ATGATTAAGT TACGTATTGC CAACGCTATC GGTATTGACA CTTCTGGTAT TCTACTCACC GAAGAACAGA
       AGCAACAGAA GATGGCCCAA CAGTCTATGC AAATGGGTAT GGATAATGGT GCTGCTGCGC TGGCTCAAGG TATGGCTGCA CAAGCTACAG CTTCACCTGA
       GGCTATGGCT GCTGCCGCTG ATTCCGTAGG TTTACAGCCG GGAATTTAAT ACGACTCACT ATAGGGAGAC CTCATCTTTG AAATGAGCGA TGACAAGAGG
       TTGGAGTCCT CGGTCTTCCT GTAGTTCAAC TTTAAGGAGA CAATAATAAT GGCTGAATCT AATGCAGACG TATATGCATC TTTTGGCGTG AACTCCGCTG
22000  TGATGTCTGG TGGTTCCGTT GAGGAACATG AGCAGAACAT GCTGGCTCTT GATGTTGCTG CCCGTGATGG CGATGATGCA ATCGAGTTAG CGTCAGACGA
       AGTGGAAACA GAACGTGACC TGTATGACAA CTCTGACCCG TTCGGTCAAG AGGATGACGA AGCCGCATT CAGGTTCGTA TCGGTGATGG CTCTGAGCCG
       ACCGATGTGG ACACTGGAGA AGAAGCGTT GAGGCACCG AAGGTTCCGA AGAGTTTACC CCACTGGGCG AGACTCCAGA AGAACTGGTA GCTGCCTCTG
       AGCAACTTGG TGAGCACGAA GAGGCTTCC AAGACATGAT TAACATTGCT GCTGAGCGTG GCATGAGTGT CGAGACCATT GAGGCTATCC AGCGTGAGTA
       CGAGGAGAAC GAAGAGTTGT CCGCCGAGTC CTACGCTAAG CTGGCTGAAA TTGGCTACAC GAAGGCTTTC ATTGACTCGT ATATCCGTGG TCAAGAAGCT
22500  CTGGTGGAGC AGTACGTAAA CAGTGTCATT GAGTACGCTG GTGGTCGTGA ACGTTTTGAT GCACTGTATA ACCACCTTGA GACGCACAAC CCTGAGGCTG
       CACAGTCGCT GGATAATGCG TTGACCAATC GTGACTTAGC GACCGTTAAG GCTATCATCA ACTTGGCTGG TGAGTCTCGC GCTAAGGCGT TCGGTCGTAA
       GCCAACTCGT AGTGTGACTA ATCGTGCTAT TCCGGCTAAA CCTCAGGCTA CCAAGCGTGA AGGCTTTGCG GACCGTAGCG AGATGATTAA AGCTATGAGT
       GACCCTCGGT ATCGCACAGA TGCCAACTAT CGTCGTCAAG TCGAACAGAA AGTAATCGAT TCGAACTTCT GATAGACTTC GAAATTAATA CGACTCACTA
       TAGGGAGACC ACAACGGTTT CCCTCTAGAA ATAATTTTGT TTAACTTTAA GAAGGAGATA TACATATGGC TAGCATGACT GGTGGACAGC AAATGGGTAC
23000  TAACCAAGGT AAAGGTGTAG TTGCTGCTGG AGATAAACTG GCGTTGTTCT TGAAGGTATT TGGCGGTGAA GTCCTGACTG CGTTCGCTCG TACCTCGTG
       ACCACTTCTC GCCACATGGT ACGTTCCATC TCCAGCGGTA AATCCGCTCA GTTCCCTGTT CTGGGTCGCA CTCAGGCAGC GTATCTGGCT CCGGGCGAGA
       ACCTCGACGA TAAACGTAAG GACATCAAAC ACACCGAGAA GGTAATCACC ATTGACGGTC TCCTGACGGC TGACGTTCTG ATTTATGATA TTGAGGACGC
       GATGAACCAC TACGACGTTC GCTCTGAGTA TACCTCTCAG TTGGGTGAAT CTCTGGCGAT GGCTGCGGAT GGTGCGGTTC TGGCTGAGAT TGCCGGTCTG
       TGTAACGTGG AAAGCAAATA TAATGAGAAC ATCGAGGGCT TAGGTACTGC TACCGTAATT GAGACCACTC AGAACAAGGC CGCACTTACC GACCAAGTTG
23500  CGCTGGGTAA GGAGATTATT GCGGCTCTGA CTAAGGCTCG TGCGGCTCTG ACCAAGAACT ATGTTCCGGC TGCTGACCGT GTGTTCTACT GTGACCCAGA
       TAGCTACTCT GCGATTCTGG CAGCACTGAT GCCGAACGCA GCAAACTACG CTGCTCTGAT TGACCCTGAG AAGGGTTCTA TCCGCAACGT TATGGGCTTT
       GAGGTTGTAG AAGTTCCGCA CCTCACCGCT GGTGGTGCTG GTACCGCTCG TGAGGGCACT ACTGGTCAGA AGCACGTCTT CCCTGCCAAT AAAGGTGAGG
       GTAATGTCAA GGTTGCTAAG GACAACGTTA TCGGCCTGTT CATGCACCGC TCTGCGGTAG GTACTGTTAA GCTGCGTGAC TTGGCTCTGG AGCGCGCTCG
       CCGTGCTAAC TTCCAAGCGC ACCAGATTAT CGCTAAGTAC GCAATGGGCC ACGGTGGTCT TCGCCCAGAA GCTGCTGGTG CAGTGGTTTT CAAAGTGGAG
24000  TAATGCTGGG GGTGGCCTCA ACGGTCGCTG CTAGTCCCGA AGAGGCGAGT GTTACTTCAA CAGAAGAAAC CTTAACGCCA GCACAGGAGG CCGCACGCAC
       CCGCGCTGCT AACAAAGCCC GAAAGGAAGC TGAGTTGGCT GCTGCCACCG CTGAGCAATA ACTAGCATAA CCCCTTGGGG CCTCTAAACG GGTCTTGAGG
       GGTTTTTTGC TGAAAGGAGG AACTATATGC GCTCATACGA TATGAACGTT GAGACTGCCG CTGAGTTATC AGCTGTCAAC GACATTCTGG CGTCTATCGG
       TGAACCTCCG GTATCAACGC TGGAAGGTGA CGCTAACGCA GATGCAGCGA ACGCTCGCCG TATTCTCAAC AAGATTAACC GACAGATTCA ATCTCGTGGA
       TGGACGTTCA ACATTGAGGA AGGCATAACG CTACTACCTG ATGTTTACTC CAACCTGATT GTATACAGTG ACGACTATTT ATCCCTAATG TCTACTTCCG
24500  GTCAATCCAT CTACGTTAAC CGAGGTGGCT ATGTGTATGA CCGAACGAGT CAATCAGACC GCTTTGACTC TGGTATTACT GTGAACATTA TTCGTCTCCG
       CGACTACGAT GAGATGCCTG AGTGCTTCCG TTACTGGATT GTCACCAAGG CTTCCCGTCA GTTCAACAAC CGATTCTTTG CGGCACCGGA AGTACAGGGT
       GTACTCCAAG AAGAGGAAGA TGAGGCTAGA CGTCTCTGCA TGGAGTATGA GATGCACTAC GGTGGGTACA ATATGCTGGA TGGAGATGCG TTCACTTCTG
       GTCTACTGAC TCGCTAACAT TAATAAATAA GGAGGCTCTA ATGGCACTCA TTAGCCAATC AATCAAGAAC TTGAAGGGTG GTATCAGCCA ACAGCCTGAC
       ATCCTTCGTT ATCCAGACCA AGGGTCACGC CAAGTTAACG GTTGGTCTTC GGAGACCGAG GGCCTCCAAA AGCGTCCACC TCTTGTTTTC TTAAATACAC
```

FIG. 1(e)

```
25000  TTGGAGACAA CGGTGCGTTA GGTCAAGCTC CGTACATCCA CCTGATTAAC CGAGATGAGC ACGAACAGTA TTACGCTGTG TTCACTGGTA GCGGAATCCG
       AGTGTTCGAC CTTTCTGGTA ACGAGAAGCA AGTTAGGTAT CCTAACGGTT CCAACTACAT CAAGACCGCT AATCCACGTA ACGACCTGCG AATGGTTACT
       GTAGCAGACT ATACGTTCAT CGTTAACCGT AACGTTGTTG CACAGAAGAA CACAAAGTCT GTCAACTTAC CGAATTACAA CCCTAATCAA GACGGATTGA
       TTAACGTTCG TGGTGGTCAG TATGGTAGGG AACTAATTGT ACACATTAAC GGTAAAGACG TTGCGAAGTA TAAGATACCA GATGGTAGTC AACCTGAACA
       CGTAAACAAT ACGGATGCCC AATGGTTAGC TGAAGAGTTA GCCAAGCAGA TCGCACTAA CTTGTCTGAT TGGACTGTAA ATGTAGGGCA AGGGTTCATC
25500  CATGTGACCG CACCTAGTGG TCAACAGATT GACTCCTTCA CGACTAAAGA TGGCTACGCA GACCAGTTGA TTAACCCTGT GACCCACTAC GCTCAGTCGT
       TCTCTAAGCT GCCACCTAAT GCTCCTAACG GCTACATGGT GAAAATCGTA GGGGACGCCT CTAAGTCTGC CGACCAGTAT TACGTTCGGT ATGACGCTGA
       GCGGAAAGTT TGGACTGAGA CTTTAGGTTG GAACACTGAG GACCAAGTTC TATGGGAAAC CATGCCACAC GCTCTTGTGC GAGCCGCTGA CGGTAATTTC
       GACTTCAAGT GGCTTGAGTG GTCTCCTAAG TCTTGTGGTG ACGTTGACAC CAACCCTTGG CCTTCTTTTG TTGGTTCAAG TATTAACGAT GTGTTCTTCT
       TCCGTAACCG CTTAGGATTC CTTAGTGGGG AGAACATCAT ATTGAGTCGT ACAGCCAAAT ACTTCAACTT CTACCCTGCG TCCATTGCGA ACCTTAGTGA
26000  TGACGACCCT ATAGACGTAG CTGTGAGTAC CAACCGAATA GCAATCCTTA AGTACGCCGT TCCGTTCTCA GAAGAGTTAC TCATCTGGTC CGATGAAGCA
       CAATTCGTCC TGACTGCCTC GGGTACTCTC ACATCTAAGT CGGTTGAGTT GAACCTAACG ACCCAGTTTG ACGTACAGGA CCGAGCGAGA CCTTTTGGGA
       TTGGGCGTAA TGTCTACTTT GCTAGTCCGA GGTCCAGCTT CACGTCCATC CACAGGTACT ACGCTGTGCA GGATGTCAGT TCCGTTAAGA ATGCTGAGGA
       CATTACATCA CACGTTCCTA ACTACATCCC TAATGGTGTG TTCAGTATTT GCGGAAGTGG TACGGAAAAC TTCTGTTCGG TACTATCTCA CGGGGACCCT
       AGTAAAATCT TCATGTACAA ATTCCTGTAC CTGAACGAAG AGTTAAGGCA ACAGTCGTGG TCTCATTGGG ACTTTGGGGA AAACGTACAG GTTCTAGCTT
26500  GTCAGAGTAT CAGCTCAGAT ATGTATGTGA TTCTTCGCAA TGAGTTCAAT ACGTTCCTAG CTAGAATCTC TTTCACTAAG AACGCCATTG ACTTACAGGG
       AGAACCCTAT CGTGCCTTTA TGGACATGAA GATTCGATAC ACGATTCCTA GTGGAACATA CAACGATGAC ACATTCACTA CCTCTATTCA TATTCCAACA
       ATTTATGGTG CAAACTTCGG GAGGGGCAAA ATCACTGTAT TGGAGCCTGA TGGTAAGATA ACCGTGTTTG ACGAACCTAC GGCTGGGTGG AATAGCGACC
       CTTGGCTGAG ACTCAGCGGT AACTTGGAGG GACGCATGGT GTACATTGGG TTCAACATTA ACTTCGTATA TGAGTTCTCT AAGTTCCTCA TCAAGCAGAC
       TGCCGACGAC GGGTCTACCT CCACGGAAGA CATTGGGCGC TTACAGTTAC GCCGAGCGTG GGTTAACTAC GAGAACTCTG GTACGTTTGA CATTTATGTT
27000  GAGAACCAAT CGTCTAACTG GAAGTACACA ATGGCTGGTG CCCGATTAGG CTCTAACACT CTGAGGGCTG GGAGACTGAA CTTAGGGACC GGACAATATC
       GATTCCCTGT GGTTGGTAAC GCCAAGTTCA ACACTGTATA CATCTTGTCA GATGAGACTA CCCCTCTGAA CATCATTGGG TGTGGCTGGG AAGGTAACTA
       CTTACGGAGA AGTTCCGGTA TTTAATTAAA TATTCTCCCT GTGGTGGCTC GAAATTAATA CGACTCACTA TAGGGAGAAC AATACGACTA CGGGAGGGTT
       TTCTTATGAT GACTATAAGA CCTACTAAAA GTACAGACTT TGAGGTATTC ACTCCGGCTC ACCATGACAT TCTTGAAGCT AAGGCTGCTG GTATTGAGCC
       GAGTTTCCCT GATGCTTCCG AGTGTGTCAC GTTGAGCCTC TATGGGTTCC CTCTAGCTAT CGGTGGTAAC TGCGGGGACC AGTGCTGGTT CGTTACGAGC
27500  GACCAAGTGT GGCGACTTAG TGGAAAGGCT AAGCGAAAGT TCCGTAAGTT AATCATGGAG TATCGCGATA AGATGCTTGA GAAGTATGAT ACTCTTTGGA
       ATTACGTATG GGTACGCAAT ACGTCCCACA TTCGTTTCCT CAAGACTATC GGTGCGGTAT TCCATGAAGA GTACACACGA GATGGTCAAT TTCAGTTATT
       TACAATCACG AAAGGAGGAT AACCATATGT GTTGGGCAGC CGCAATACCT ATCGCTATAT CTGGCGCTCA GGCTATCAGT GGTCAGAACG CTCAGGCCAA
       AATGATTGCC GCTCAGACCG CTGCTGGTCG TCGTCAAGCT ATGGAAATCA TGAGGCAGAC GAACATCCAG AATGCTGACC TATCGTTGCA AGCTCGAAGT
       AAACTTGAGG AAGCGTCCGC CGAGTTGACC TCACAGAACA TGCAGAAGGT CCAAGCTATT GGGTCTATCC GAGCGGCTAT CGGAGAGAGT ATGCTTGAAG
28000  GTTCCTCAAT GGACCGCATT AAGCGAGTCA CAGAAGGACA GTTCATTCGG GAAGCCAATA TGGTAACTGA GAACTATCGC CGTGACTACC AAGCAATCTT
       CGCACAGCAA CTTGGTGGTA CTCAAAGTGC TGCAAGTCAG ATTGACGAAA TCTATAAGAG CGAACAGAAA CAGAAGAGTA AGCTACAGAT GGTTCTGGAC
       CCACTGGCTA TCATCGGGTC TTCCGCTGCG AGTGCTTACG CATCCGGTGC GTTCGACTCT AAGTCCACAA CTAAGGCACC TATTGTTGCC GCTAAAGGAA
       CCAAGACGGG GAGGTAATGA GCTATGAGTA AAATTGAATC TGCCCTTCAA GCGGCACAAC CGGGACTCTC TCGGTTACGT GGTGGTGCTG GAGGTATGGG
       CTATCGTGCA GCAACCACTC AGGCCGAACA GCCAAGGTCA AGCCTATTGG ACACCATTGG TCGGTTCGCT AAGGCTGGTG CCGATATGTA TACCGCTAAG
28500  GAACAACGAG CACGAGACCT AGCTGATGAA CGCTCTAACG AGATTATCCG TAAGCTGACC CCTGAGCAAC GTCGAGAAGC TCTCAACAAC GGGACCCTTC
       TGTATCAGGA TGACCCATAC GCTATGGAAG CACTCCGAGT CAAGACTGGT CGTAACGCTG CGTATCTTGT GGACGATGAC GTTATGCAGA AGATAAAAGA
       GGGTGTCTTC CGTACTCGCG AAGAGATGGA AGAGTATCGC CATAGTCGCC TTCAAGAGGG CGCTAAGGTA TACGCTGAGC AGTTCGGCAT CGACCCTGAG
       GACGTTGATT ATCAGCGTGG TTTCAACGGG GACATTACCG AGCGTAACAT CTCGCTGTAT GGTGCGCATG ATAACTTCTT GAGCCAGCAA GCTCAGAAGG
       GCGCTATCAT GAACAGCCGA GTGGAACTCA ACGGTGTCCT TCAAGACCCT GATATGCTGC GTCGTCCAGA CTCTGCTGAC TTCTTTGAGA AGTATATCGA
29000  CAACGGTCTG GTTACTGGCG CAATCCCATC TGATGCTCAA GCCACACAGC TTATAAGCCA AGCGTTCAGT GACGCTTCTA GCCGTGCTGG TGGTGCTGAC
       TTCCTGATGC GAGTCGGTGA CAAGAAGGTA ACACTTAACG GAGCCACTAC GACTTACCGA GAGTTGATTG GTGAGGAACA GTGGAACGCT CTCATGGTCA
       CAGCACAACG TTCTCAGTTT GAGACTGACG CGAAGCTGAA CGAGCAGTAT CGCTTGAAGA TTAACTCTGC GCTGAACCAA GAGGACCCAA GGACAGCTTG
       GGAGATGCTT CAAGGTATCA AGGCTGAACT AGATAAGGTC CAACCTGATG AGCAGATGAC ACCACAACGT GAGTGGCTAA TCTCCGCACA GGAACAAGTT
       CAGAATCAGA TGAACGCATG GACGAAAGCT CAGGCCAAGG CTCTGGACGA TTCCATGAAG TCAATGAACA AACTTGACGT AATCGACAAG CAATTCCAGA
29500  AGCGAATCAA CGGTGAGTGG GTCTCAACGG ATTTTAAGGA TATGCCAGTC AACGAGAACA CTGGTGAGTT CAAGCATAGC GATATGGTTA ACTACGCCAA
       TAAGAAGCTC GCTGAGATTG ACAGTATGGA CATTCCAGAC GGTGCCAAGG ATGCTATGAA GTTGAAGTAC CTTCAAGCGG ACTCTAAGGA CGGAGCATTC
       CGTACAGCCA TCGGAACCAT GGTCACTGAC GCTGGTCAAG AGTGGTCTGC CGCTGTGATT AACGGTAAGT TACCAGAACG AACCCCAGCT ATGGATGCTC
       TGCGCAGAAT CCGCAATGCT GACCCTCAGT TGATTCTGC GCTATACCCA GACCAAGCTG AGCTATTCCT GACGATGGAC ATGATGGACA AGCAGGGTAT
       TGACCCTCAG GTTATTCTTG ATGCCGACCG ACTGACTGTT AAGCGGTCCA AAGAGCAACG CTTTGAGGAT GATAAGCAT TCGAGTCTGC ACTGAATGCA
```

FIG. 1(f)

30000 TCTAAGGCTC CTGAGATTGC CCGTATGCCA GCGTCACTGC GCGAATCTGC ACGTAAGATT TATGACTCCG TTAAGTATCG CTCGGGGAAC GAAAGCATGG
CTATGGAGCA GATGACCAAG TTCCTTAAGG AATCTACCTA CACGTTCACT GGTGATGATG TTGACGGTGA TACCGTTGGT GTGATTCCTA AGAATATGAT
GCAGGTTAAC TCTGACCCGA AATCATGGGA GCAAGGTCGG GATATTCTGG AGGAAGCACG TAAGGGAATC ATTGCGAGCA ACCCTTGGAT AACCAATAAG
CAACTGACCA TGTATTCTGA AGGTGACTCC ATTTACCTTA TGGACACCAC AGGTCAAGTC AGAGTCCGAT ACGACAAAGA GTTACTCTCG AAGGTCTGGA
GTGAGAACCA GAAGAAACTC GAAGAGAAAG CTCGTGAGAA GGCTCTGGCT GATGTGAACA AGCGAGCACC TATAGTTGCC GCTACGAAGG CCGTGAAGC
30500 TGCTGCTAAA CGAGTCCGAG AGAAACGTAA ACAGACTCCT AAGTTCATCT ACGGACGTAA GGAGTAACTA AAGGCTACAT AAGGAGGCCC TAAATGGATA
AGTACGATAA GAACGTACGA AGTGATTATG ATGGTCTGTT CCAAAAGGCT GCTGATGCCA ACGGGGTCTC TTATGACCTT TTACGTAAAG TCGCTTGGAC
AGAATCACGA TTTGTGCCTA CAGCAAAATC TAAGACTGGA CCATTAGGCA TGATGCAATT TACCAAGGCA ACCGCTAAGG CCCTCGGTCT GCGAGTTACC
GATGGTCCAG ACGACGACCG ACTGAACCCT GAGTTAGCTA TTAATGCTGC CGCTAAGCAA CTTGCAGGTC TGGTAGCGAA GTTTGATGGC GATGAACTCA
AAGCTGCCCT TGCGTACAAC CAAGGCGAGG GACGCTTGGG TAATCCACAA CTTGAGGCGT ACTCTAAGGG AGACTTCGCA TCAATCTCTG AGGATGGACG
31000 TAACTACATG CGTAACCTTC TGGATGTTGC TAAGTCACCT ATGGCTGGAC AGTTGGAAAC TTTTGGTGGC ATAACCCCAA AGGGTAAAGG CATTCGGGCT
GAGGTAGGAT TGGCTGGAAT TGGTCACAAG CAGAAAGTAA CACAGGAACT TCCTGAGTCC ACAAGTTTTG ACGTTAAGGG TATCGAACAG GAGGCTACGG
CGAAACCATT CGCCAAGGAC TTTTGGGAGA CCCACGGAGA AACACTTGAC GAGTACAACA GTCGTTCAAC CTTCTTCGGA TTCAAAAATG CTGCCGAAGC
TGAACTCTCC AACTCAGTCG CTGGGATGGC TTTCCGTGCT GGTCGTCTCG ATAATGGTTT TGATGTGTTT AAAGACAACA TTACGCCGAC TCGCTGGAAC
TCTCACATCT GGACTCCAGA GGAGTTAGAG AAGATTCGAA CAGAGGTTAA GAACCCTGCG TACATCAAGG TTGTAACTGG TGGTTCCGCT GAGAACCTCG
31500 ATGACCTCAT TAAATTGGCT AACGAGAACT TTGAGAATGA CTCCCGCGCT GCCGAGGCTG GCCTAGGTGC CAAACTGAGT GCTGGTATTA TTGGTGCTGG
TGTGGACCCG CTTAGCTATG TTCCTATGGT CGGTGTCACT GGTAAGGGCT TTAAGTTAAT CAATAAGGCT CTTGTAGTTG GTGCCGAAAG TGCTGCTCTG
AACGTTGCAT CCGAAGGTCT CCGTACCTCC GTAGCTGGTG GTGACGCAGA CTATGCGGGT GCTGCCTTAG GTGGCTTTGT GTTTGGCGCA GGCATGTCTG
CAATCAGTGA CGCTGTAGCT GCTGGACTGA AACGCACTAA ACCAGAAGCT GAGTTCGACA ATGAGTTCAT CGGTCCTATC ATGCGATTGG AAGCCTGTGA
GACAGCACGA AACGCCAACT CTGCGGACCT CTCTCGGATG AACACTGAGA ACATGAAGTT TGAAGGTGAA CATAATGGTG TCCCTTATGA GGACTTACCA
32000 ACAGAGAGAG GTGCCGTGGT GTTACATGAT GGCTCCGTTC TAAGTGCAAG CAACCCAATC AACCCTAAGA CTCTAAAAGA GTTCTCCGAG GTTGACCCTG
AGAAGGCTGC GCGAGGAATC AAACTGGCTG GGTTCACCGA GATTGGCTTG AAGACCTTGG GGTCTGACGA TGCTGACATC CGTAGAGTGG CTATCGACCT
CGTTCGCTCT CCTACTGGTA TGCAGTCTGG TGCCTCAGGT AAGTTCGGTG CAACAGCTTC TGACATCCAT GAGAGACTTC ATGGTACTGA CCACCGTACT
TATAATGACT TGTACAAAGC AATGTCTGAC GCTATGAAAG ACCCTGACTT CTCTACTGGC GGCGCTAAGA TGTCCCGTGA AGAAACTCGA TACACTATCT
ACCGTAGAGC GGCACTAGCT ATTGACGCTC CAGAACTACA GAAGGCACTC ACTCCGTCTG AGAGAATCGT TATGGACATC ATTAAGCGTC ACTTTGACAC
32500 CAAGCGTGAA CTTATGGAAA ACCCACGCAT ATTCGGTAAC ACAAAGGCTG TGAGTATCTT CCCTGAGAGT CGCCACAAAG GTACTTACGT TCCTCACGTA
TATGACCGTC ATGCCAAGGC GCTGATGATT CAACGCTACG GTGCCGAAGG TTTGCAGGAA GGGATTGCCC GCTCATGGAT GAACAGCTAC GTCTCCAGAC
CTGAGGTCAA GGCCAGAGTC GATGAGATGC TTAAGGAATT ACACGGGGTG AAGGAAGTAA CACCAGAGAT GGTAGAGAAG TACGCTATGG ATAAGGCTTA
TGGTATCTCC CACTCAGACC AGTTCACCAA CAGTTCCATA ATAGAAGAGA ACATTGAGGG CTTAGTAGGT ATCGAGAATA ACTCATTCCT TGAGGCACGT
AACTTGTTTG ATTCGGACCT ATCCATCACT ATGCCAGACG GACAGCAATT CTCAGTGAAT GACCTAAGGG ACTTCGATAT GTTCCGCATC ATGCCAGCGT
33000 ATGACCGCCG TGTCAATGGT GACATCGCCA TCATGGGGTC TACTGGTAAA ACCACTAAGG AACTTAAGGA TGAGATTTTG GCTCTCAAAG CGAAAGCTGA
GGGAGACGGT AAGAAGACTG GCGAGGTACA TGCTTTAATG GATACCGTTA AGATTCTTAC TGGTCGTGCT AGACGCAATC AGGACACTGT GTGGGAAACC
TCACTGCGTG CCATCAATGA CCTAGGGTTC TTCGCTAAGA ACGCCTACAT GGGTGCTCAG AACATTACGG AGATTGCTGG GATGATTGTC ACTGGTAACG
TTCGTGCTCT AGGGCATGGT ATCCCAATTC TGCGTGATAC ACTCTACAAG TCTAAACCAG TTTCAGCTAA GGAACTCAAG GAACTCCATG CGTCTCTGTT
CGGCAAGGAG GTGGACCAGT TGATTCGGCC TAAACGTGCT GACATTGTGC AGCGCCTAAG GGAAGCAACT GATACCGGAC CTGCCGTGGC GAACATCGTA
33500 GGGACCTTGA AGTATTCAAC ACAGGAACTG GCTGCTCGCT CTCCGTGGAC TAAGCTACTG AACGGAACCA CTAACTACCT TCTGGATGCT GCGCGTCAAG
GTATGCTTGC GGATGTTATT AGTGCCACCC TAACAGGTAA GACTACCCGC TGGGAGAAAG AAGGCTTCCT TCGTGGTGCC TCCGTAACTC CTGAGCAGAT
GGCTGGCATC AAGTCTCTCA TCAAGGAACA TATGGTACGC GGTGAGGACG GGAAGTTTAC CGTTAAGGAC AAGCAAGCGT TCTCTATGGA CCCACGGGCT
ATGGACTTAT GGAGACTGGC TGACAAGGTA GCTGATGAGG CAATGCTGCG TCCACATAAG GTGTCCTTAC AGGATTCCCA TGCGTTCGGA GCACTGGTA
AGATGGTTAT GCAGTTTAAG TCTTTCACTA TCAAGTCCCT TAACTCTAAG TTCCTGCGAA CCTTCTATGA TGGATACAAG AACAACCGAG CGATTGACGC
34000 TGCGCTGAGC ATCATCACCT CTATGGGTCT CGCTGGTGGT TTCTATGCTA TGGCTGCACA CGTCAAAGCA TACGCTCTGC CTAAGGAGAA ACGTAAGGAG
TACTTGGAGC GTGCACTGGA CCCAACCATG ATTGCCCACG CTGCGTTATC TCGTAGTTCT CAATTGGGTG CTCCTTTGGC TATGGTTGAC CTAGTTGGTG
GTGTTTTAGG GTTCGAGTCC TCCAAGATGG CTCGCTCTAC GATTCTACCT AAGGACACCG TGAAGGAACG TGACCCAAAC AAACCGTACA CCTCTAGACA
GGTAATGGGC GCTATGGGTT CAAACCTTCT GGAACAGATG CCTTCGGCTG GCTTTGTGGC TAACGTAGGG GCTACCTTAA TGAATGCTGC TGGCGTGGTG
AACTCACCTA ATAAAGCAAC CGAGCAGGAC TTCATGACTG GTCTTATGAA CTCCACAAAA GAGTTAGTAC CGAACGACCC ATTGACTCAA CAGCTTGTG
34500 TGAAGATTTA TGAGCGAAC GGTGTTAACT TGAGGGACGG TACGAAATAA TACGACTCAC TATAGGGAGA GGCGAAATAA TCTTCTCCCT GTAGTCTCTT
AGATTTACTT TAAGCAGGTC AAATGGCTAA CGTAATTAAA ACCGTTTTGA CTTACCAGTT AGATGGCTCC AATCGTGATT TTAATATCCC GTTTGAGTAT
CTAGCCCGTA AGTTCGTAGT GGTAACTCTT ATTGGTGTAG ACCGAAAGGT CCTTACGATT AATACAGACT ATCGCTTTGC TACACGTACT ACTATCTATG
TGACAAAGGC TTGGGGTCCA GCCGATGGCT ACACGACCAT CGAGTTACGT CGAGTAACCT CCACTACCGA CCGATTGGTT GACTTTACGG ATGGTTCAAT
CCTCCGCGCG TATGACCTTA ACGTCGCTCA GATTCAAACG ATGCACGTAG CGGAACAGGC CCGTGACCTC ACTACGGATA CTATCGGTGT CAATAACGAT

FIG. 1(g)

```
35000  GGTCACTTGG ATGCTCGTGG TCGTCGAATT GTGAACCTAG CGAACGCCGT GGATGACCGC GATGCTGTTC CGTTTGGTCA ACTAAAGACC ATGAACCAGA
       ACTCATGGCA AGCACGTAAT GAAGCCTTAC AGTTCCGTAA TGAGGCTGAG ACTTTCAGAA ACCAAGCGGA GGGCTTTAAG AACGAGTCCA GTACCAACGC
       TACGAACACA AAGCAGTGGC GCGATGAGAC CAAGGGTTTC CGAGACGAAG CCAAGCGGTT CAAGAATACG GCTGGTCAAT ACGCTACATC TGCTGGGAAC
       TCTGCTTCCG CTGCGCATCA ATCTGAGGTA AACGCTGAGA ACTCTGCCAC AGCATCCGCT AACTCTGCTC ATTTGGCAGA ACAGCAAGCA GACCGTGCGG
       AACGTGAGGC AGACAAGCTG GAAAATTACA ATCGATTGGC TGGTGCAATT GATAAGGTAG ATGGAACCAA TGTGTACTGC AAACGAAATA TTCACGCTAA
35500  CGGGCGCCTT TACATGACCA CAAACGGTTT TGACTGTGGC CAGTATCAAC AGTTCTTTGG TGGTGTCACT AATCGTTACT CTGTCATGGA GTGGGGAGAT
       GAGAACGGAT GGCTGATGTA TGTTCAACGT AGAGAGTGGA CAACAGCGAT AGGCGGTAAC ATCCAGTTAG TAGTAAACGG ACAGATCATC ACCCAAGGTG
       GAGCCATGAC CGGTCAGCTA AAATTGCAGA ATGGGCATGT TCTTCAATTA GAGTCCGCAT CCGACAAGGC GCACTATATT CTATCTAAAG ATGGTAACAG
       GAATAACTGG TACATTGGTA GAGGGTCAGA TAACAACAAT GACTGTACCT TCCACTCCTA TGTACATGGT ACGACCTTAA CACTCAAGCA GGACTATGCA
       GTAGTTAACA AACACTTCCA CGTAGGTCAG GCCGTTGTGG CCACTGATGG TAATATTCAA GGTACTAAGT GGGGAGGTAA ATGGCTGGAT GCTTACCTAC
36000  GTGACAGCTT CGTTGCGAAG TCCAAGGCGT GGACTCAGGT GTGGTCTGGT AGTGCTGGCG GTGGGGTAAG TGTGACTGTT TCACAGGATC TCCGCTTCCG
       CAATATCTGG ATTAAGTGTG CCAACAACTC TTGCAACTTC TTCCGTACTG GCCCCGATGG AATCTACTTC ATAGCCTCTG ATGGTGGATG GTTACGATTC
       CAAATACACT CCAACGGTCT CGGATTCAAG AATATTGCAG ACAGTCGTTC AGTACCTAAT GCAATCATGG TGGAGAACGA GTAATTGGTA AATCACAAGG
       AAAGACGTGT AGTCCACGGA TGGACTCTCA AGGAGGTACA AGGTGCTATC ATTAGACTTT AACAACGAAT TGATTAAGGC TGCTCCAATT GTTGGGACGG
       GTGTAGCAGA TGTTAGTGCT CGACTGTTCT TTGGTTAAG CCTTAACGAA TGGTTCTACG TTGCTGCTAT CGCCTACACA GTGGTTCAGA TTGGTGCCAA
36500  CGTAGTCGAT AAGATGATTG ACTGGAAGAA AGCCAATAAG GAGTGATATG TATGGAAAAG GATAAGAGCC TTATTACATT CTTAGAGATG TTGGACACTG
       CGATGGCTCA GCGTATGCTT GCGGACCTTT CGGACCATGA GCGTCGCTCT CCGCAACTCT ATAATGCTAT TAACAAACTG TTAGACCGCC ACAAGTTCCA
       GATTGGTAAG TTGCAGCCGG ATGTTCACAT CTTAGGTGGC CTTGCTGGTG CTCTTGAAGA GTACAAAGAG AAAGTCGGTG ATAACGGTCT TACGGATGAT
       GATATTTACA CATTACAGTG ATATACTCAA GGCCACTACA GATAGTGGTC TTTATGGATG TCATTGTCTA TACGAGATGC TCCTACGTGA AATCTGAAAG
       TTAACGGGAG GCATTATGCT AGAATTTTTA CGTAAGCTAA TCCCTTGGGT TCTCGCTGGG ATGCTATTCG GGTTAGGATG GCATCTAGGG TCAGACTCAA
37000  TGGACGCTAA ATGGAAACAG GAGGTACACA ATGAGTACGT TAAGAGAGTT GAGGCTGCGA AGAGCACTCA AAGAGCAATC GATGCCGTAT CTGCTAAGTA
       TCAAGAAGAC CTTGCCGCGC TGGAAGGGAG CACTGATAGG ATTATTTCTG ATTTGCGTAG CGACAATAAG CGGTTGCGCG TCAGAGTCAA AACTACCGGA
       ACCTCCGATG GTCAGTGTGG ATTCGAGCCT GATGGTCGAG CCGAACTTGA CGACCGAGAT GCTAAACGTA TTCTCGCAGT GACCCAGAAG GGTGACGCAT
       GGATTCGTGC GTTACAGGAT ACTATTCGTG AACTGCAACG TAAGTAGGAA ATCAAGTAAG GAGGCAATGT GTCTACTCAA TCCAATCGTA ATGCGCTCGT
       AGTGGCGCAA CTGAAAGGAG ACTTCGTGGC GTTCCTATTC GTCTTATGGA AGGCGCTAAA CCTACCGGTG CCCACTAAGT GTCAGATTGA CATGGCTAAG
37500  GTGCTGGCGA ATGGAGACAA CAAGAAGTTC ATCTTACAGG CTTTCCGTGG TATCGGTAAG TCGTTCATCA CATGTGCGTT CGTTGTGTGG TCCTTATGGA
       GAGACCCTCA GTTGAAGATA CTTATCGTAT CAGCCTCTAA GGAGCGTGCA GACGCTAACT CCATCTTTAT TAAGAACATC ATTGACCTGC TGCCATTCCT
       ATCTGAGTTA AAGCCAAGAC CCGGACAGCG TGACTCGGTA ATCAGCTTTG ATGTAGGCCC AGCCAATCCT GACCACTCTC CTAGTGTGAA ATCAGTAGGT
       ATCACTGGTC AGTTAACTGG TAGCCGTGCT GACATTATCA TTGCGGATGA CGTTGAGATT CCGTCTAACA GCGCAACTAT GGGTGCCCGT GAGAAGCTAT
       GGACTCTGGT TCAGGAGTTC GCTGCGTTAC TTAAACCGCT GCCTTCCTCT CGCGTTATCT ACCTTGGTAC ACCTCAGACA GAGATGACTC TCTATAAGGA
38000  ACTTGAGGAT AACCGTGGGT ACACAACCAT TATCTGGCCT GCTCTGTACC CAAGGACACG TGAAGAGAAC CTCTATTACT CACAGCGTCT TGCTCCTATG
       TTACGCGCTG AGTACGATGA GAACCCTGAG GCACTTGCTG GGACTCCAAC AGACCCAGTG CGCTTTGACC GTGATGACCT GCGCGAGCGT GAGTTGGAAT
       ACGGTAAGGC TGCCTTTACG CTACAGTTCA TGCTTAACCC TAACCTTAGT GATGCCGAGA AGTACCCGCT GAGGCTTCGT GACGCTATCG TAGCGGCCTT
       AGACTTAGAG AAGGCCCCAA TGCATTACCA GTGGCTTCCG AACCGTCAGA ACATCATTGA GGACCTTCCT AACGTTGGCC TTAAGGGTGA TGACCTGCAT
       ACGTACCACG ATTGTTCCAA CAACTCAGGT CAGTACCAAC AGAAGATTCT GGTCATTGAC CCTAGTGGTC GCGGTAAGGA CGAAACAGGT TACGCTGTGC
38500  TGTACACACT GAACGGTTAC ATCTACCTTA TGGAAGCTGG AGGTTTCCGT GATGGCTACT CCGATAAGAC CCTTGAGTTA CTCGCTAAGA AGGCAAAGCA
       ATGGGGAGTC CAGACGGTTG TCTACGAGAG TAACTTCGGT GACGGTATGT TCGGTAAGGT ATTCAGTCCT ATCCTTCTTA AACACCACAA CTGTGCGATG
       GAAGACATTC GTGCCCGTGG TATGAAAGAG ATGCGTATTT GCGATACCCT TGAGCCAGTC ATGCAGACTC ACCGCCTTGT AATTCGTGAT GAGGTCATTA
       GGGCCGACTA CCAGTCCGCT CGTGACGTAG ACGGTAAGCA TGACGTTAAG TACTCGTTGT TCTACCAGAT GACCCGTATC ACTCGTGAGA AAGGCGCTCT
       GGCTCATGAT GACCGATTGG ATGCCTTGC GTTAGGCATT GAGTATCTCC GTGAGTCCAT GCAGTTGGAT TCCGTTAAGG TCGAGGGTGA AGTACTTGCT
39000  GACTTCTTTG AGGAACACAT GATGCGTCCT ACGGTTGCTG CTACGCATAT CATTGAGATG TCTGTGGGAG GAGTTGATGT GTACTCTGAG GACGATGAGG
       GTTACGGTAC GTCTTTCATT GAGTGGTGAT TTATGCATTA GGACTGCATA GGGATGCACT ATAGACCACG GATGGTCAGT TCTTTAAGTT ACTGAAAAGA
       CACGATAAAT TAATACGACT CACTATAGGG AGAGGAGGGA CGAAAGGTTA CTATATACAT ACTGAATGAA TACTTATAGA GTGCATAAAG TATGCATAAT
       GGTGTACCTA GAGTGACCTC TAAGAATGGT GATTATATTG TATTAGTATC ACCTTAACTT AAGGACCAAC ATAAAGGGAG GAGACTCATG TTCCGCTTAT
       TGTTGAACCT ACTGCGGCAT AGAGTCACCT ACCGATTTCT TGTGGTACTT TGTGCTGCCC TTGGGTACGC ATCTCTTACT GGAGACCTCA GTTCACTGGA
39500  GTCTGTCGTT TGCTCTATAC TCACTTGTAG CGATTACGGT CTTCCTGACC GACTGATGGC TCACCGAGGG ATTCAGCGGT ATGATTGCAT CACACCACTT
       CATCCCTATA GAGTCAAGTC CTAAGGTATA CCCATAAAGA GCCTCTAATG GTCTATCCTA AGGTCTATAC CTAAAGATAG GCCATCCTAT CAGTGTCACC
       TAAAGAGGGT CTTAGACAGG GCCTATGGAG TTCCTATAGG GTCCTTTAAA ATATACCATA AAAATCTGAG TGAGTATCTC ACAGTGTACG GACCTAAAGT
       TCCCCCATAG GGGGTACCTA AAGCCCAGCC AATCACCTAA AGTCAACCTT CGGTTGACCT TGAGGGTTCC CTAACGGTTG GGGATGACCC TTGGGTTTGT
       CTTTGGGTGT TACCTTGAGT GTCTCTCTGT GTCCCT
```

FIG. 1(h)

FIG. 1. (a) to (h) Nucleotide sequence of the *l* strand of T7 DNA. The sequence reads 5′ to 3′ from left to right. The *l* strand is, by definition, the strand that has its 5′ end at the genetic left end of T7 DNA; it has the same sequence as the T7 RNAs. The sequence hyphens have been omitted for clarity.

TABLE 1

Transcription and translation signals in the early region of T7 DNA

RNA polymerase[a] *E. coli*	RNA polymerase[a] T7	RNAase III[b] sites	Protein[c]	RF[d]	fMet[e] +aa	Position in T7 DNA Nucleotides	Position in T7 DNA T7 units[f]
A0						224	0·56
	ϕ0L					405	1·01
A1						498	1·25
A2						626	1·57
A3						750	1·88
		R*0·3*				890	2·23
			0·3	1	117 −	925–1276	2·32–3·20
			0·4	3	51 −	1278–1431	3·20–3·58
		R*0·5*				1468	3·68
			0·5	2	47 +	1496–1637	3·75–4·10
			0·6A	1	53 +	1636–1795	4·10–4·49
			0·6B	1·2	111 +	1636–1970	4·10–4·93
			0·7	2	359 +	2021–3098	5·06–7·76
		R*1*				3138	7·86
			1	3	883 +	3171–5820	7·94–14·57
	*ϕ1·1*A					5848	14·64
		R*1·1*				5887	14·74
	*ϕ1·1*B					5923	14·83
			1·1	1	42 +	6007–6133	15·04–15·36
			1·2	2	85 −	6137–6392	15·37–16·01
	ϕ1·3					6409	16·05
		R*1·3*				6448	16·15
			1·3	1	359 +	6475–7552	16·21–18·91
TE						7588	19·00

Promoters, transcription termination sites, RNAase III cleavage sites, and T7 proteins are identified in the text. Maps of their relative positions are given in Figs 2 and 4.

[a] The nucleotide given for a promoter is the known or predicted first nucleotide of the RNA chain initiated at that promoter, and the nucleotide given for a transcription termination site is the last nucleotide of the majority of the chains that end at that site. The only transcription signals given for *E. coli* RNA polymerase are the 3 major early promoters (A1, A2 and A3), the minor leftward promoter near the left end (A0), and the transcription termination site at the end of the early region (TE); the positions of other minor promoters for *E. coli* RNA polymerase are given in Table 6.

[b] The nucleotide given for each RNAase III cleavage site is the known or predicted last nucleotide at the 3′ end of the RNA chain produced by cleavage at the site. Cleavage at sites R*3·5* and R*13* is relatively inefficient; this is indicated by parentheses in Tables 2 and 3.

[c] The gene numbers for the 5 potential overlapping genes are given in parentheses in Tables 2 and 3. The nucleotides given for each protein are the first nucleotides of the initiation and termination codons.

[d] The reading frame (RF) is determined by dividing the number of the first nucleotide of the initiation codon by 3; if the remainder is 1/3, the reading frame is 1; if 2/3, the reading frame is 2; if 0, the reading frame is 3. Two reading frames are shown for proteins that are thought to arise by a frameshift during translation.

[e] The number is the total amino acids (aa) predicted for the protein, including the initiating methionine. The initiating methionine is known to be removed from at least some of the proteins; proteins predicted to retain the initiating methionine are indicated by +, and those predicted to lose the initiating methionine are indicated by − (see Dunn & Studier, 1981).

[f] A T7 unit is 399·36 nucleotides or nucleotide pairs.

TABLE 2

Transcription and translation signals in the class II region of T7 DNA

RNA polymerase *E. coli*	RNA polymerase T7	RNAase III sites	Protein	RF	fMet + aa	Position in T7 DNA Nucleotides	T7 units
			1·4	3	51 +	7608–7761	19·05–19·43
	ϕ1·5					7778	19·48
			1·5	3	29 +	7791–7878	19·51–19·73
	ϕ1·6					7895	19·77
			1·6	1	86 +	7906–8164	19·80–20·44
			1·7	3	196 −	8166–8754	20·45–21·92
			1·8	1	48 +	8749–8893	21·91–22·27
			2	3	64 −	8898–9090	22·28–22·76
	ϕ2·5					9107	22·80
			2·5	2	232 −	9158–9854	22·93–24·67
			2·8	2	139 +	9857–10.274	24·68–25·73
			3	3	149 −	10.257–10.704	25·68–26·80
			3·5	2	151 −	10.706–11.159	26·81–27·94
	ϕ3·8					11.180	27·99
		(R3·8)				11.203	28·05
			3·8	2	121 +	11.225–11.588	28·11–29·02
			4·4	3	566 +	11.565–13.263	28·96–33·21
			(4·1)	1	40 −	11.635–11.755	29·13–29·43
			4B	3	503 +	11.754–13.263	29·43–33·21
	ϕ4c					12.671	31·73
			(4·2)	1	112 +	12.988–13.324	32·52–33·36
	ϕ4·3					13.341	33·41
			4·3	2	70 +	13.352–13.562	33·43–33·96
			4·5	3	89 −	13.584–13.851	34·01–34·68
		R4·7				13.892	34·79
	ϕ4·7					13.915	34·84
			4·7	1	135 +	13.927–14.332	34·87–35·89
			5	1	704 +	14.353–16.465	35·94–41·23
			5·3	1	118 +	16.483–16.837	41·27–42·16
			5·5	3	99 −	16.851–17.148	42·20–42·94
			5·7	2	69 −	17.150–17.357	42·94–43·46
			6	1	348 −	17.359–18.403	43·47–46·08
			6·3	3	37 +	18.393–18.504	46·06–46·33

See the legend and footnotes to Table 1.

restriction endonucleases are at variance with the patterns actually observed (unpublished results), whereas we have not yet detected any discrepancy between observed restriction patterns and those predicted by the sequence of Fig. 1 for any enzyme tested.

Portions of the nucleotide sequence from 12.101 to 39.936, in the regions around promoters for T7 RNA polymerase, have been published by others. The sequence given in Fig. 1 is in complete agreement with the sequence around the *ϕ13* promoter (position 68·3) published by Oakley *et al.* (1979), as corrected by Rosa (1981*b*), and with the sequences around the *ϕ4·3* and *ϕ4·7* promoters (positions 33·4 and 34·8) published by Carter & McAllister (1981). There are 21 discrepancies between the sequence given in Fig. 1 and the sequences published by Rosa (1979,1981*a*) and Rosa & Andrews (1981). Each point of discrepancy has been rechecked on our original films, and in each case the sequence in Fig. 1 seems unambiguous.

The sequence given in Fig. 1 is in agreement with all of the restriction activities on T7 DNA reported by McDonell *et al.* (1977), Rosenberg *et al.* (1979) and Studier *et al.* (1979).

TABLE 3

Transcription and translation signals in the class III region of T7 DNA

RNA polymerase *E. coli*	RNA polymerase T7	RNAase III sites	Protein	RF	fMet +aa	Position in T7 DNA Nucleotides	T7 units
	φ6·5						
		R*6·5*				18.544	46·43
						18.562	46·48
			6·5	1	84+	18.604–18.856	46·58–47·22
			6·7	2	88–	18.863–19.127	47·23–47·89
			7	1	133–	19.129–19.528	47·90–48·90
			7·3	1	99–	19.534–19.831	48·91–49·66
			7·7	2	130+	19.847–20.237	49·70–50·67
	φ9		*8*	1	536–	20.239–21.847	50·68–54·71
						21.864	54·75
	φ10		*9*	1	307–	21.949–22.870	54·96–57·27
						22.903	57·35
			10A	1	345–	22.966–24.001	57·51–60·10
	T*φ*		*10B*	1–3	398–	22.966–24.159	57·51–60·49
						24.209	60·62
			11	2	196+	24.227–24.815	60·66–62·14
	φ13		*12*	1	794–	24.841–27.223	62·20–68·17
		(R*13*)				27.273	68·29
						27.280	68·31
			13·	3	138+	27.306–27.720	68·37–69·41
			14	1	196–	27.727–28.315	69·43–70·90
			15	1	747–	28.324–30.565	70·92–76·53
	φ17		*16*	3	1318+	30.594–34.548	76·61–86·51
						34.565	86·55
			17	3	553–	34.623–36.282	86·70–90·85
			17·5	1	67+	36.343–36.544	91·00–91·51
			18	3	89+	36.552–36.819	91·53–92·20
		R*18·5*				36.855	92·29
			18·5	1	143+	36,916–37.345	92·44–93·51
			(*18·7*)	2	83–	37.031–37.280	92·73–93·35
			19	1	586–	37.369–39.127	93·57–97·97
			(*19·2*)	2	85–	38.015–38.270	95·19–95·83
	φOR		(*19·3*)	2	57–	38.552–38.723	96·53–96·96
						39.228	98·23
			19·5	1	49+	39.388–39.535	98·63–99·00

See the legend and footnotes to Table 1.

rthermore, no discrepancies have been found between the predicted and observed ctrophoretic patterns of DNA fragments produced by digestion with *Acc*I fragments). *Ava*II (55 fragments), *Fnu*4HI (157 fragments). *Hae*III (69 fragments), *ni*AI (71 fragments). *Hind*II (62 fragments). *Hpa*II (59 fragments). *Nci*I (10 fragments). ql (112 fragments). and *Tha*I (66 fragments). Restriction enzymes were obtained from thesda Research Laboratories, Inc. or from New England BioLabs.

The above digests were analyzed by electrophoresis in gels of agarose or a gradient of 3% 20% polyacrylamide, using a buffer of 40 mM-Tris, 20 mM-acetic acid, 2 mM-Na_3EDTA osenberg *et al.*, 1979). Under these conditions, relative mobility of the DNA fragments s generally a smooth function of molecular weight. However, when the same digests re analyzed by electrophoresis in polyacrylamide gels of uniform concentration, many of fragments had a relative mobility markedly different from that predicted from their ative size. Apparently, factors other than size can have a significant influence on relative

mobility of DNA fragments in uniform polyacrylamide gels, at least in this buffer system, whereas size of the fragment is the dominant determinant of relative mobility in gradient polyacrylamide gels of appropriate composition.

3. T7 Genes

(a) *Identification*

Genetic and biochemical analyses of T7 have provided evidence for the existence of at least 38 T7 genes (Hausmann & Gomez, 1967; Studier, 1969, 1972, 1973*a*,*b*, 1975*a*, 1981; Studier *et al.*, 1979; Pao & Speyer, 1975; Botstein, personal communication; unpublished results). All 38 of these known genes have been located in the nucleotide sequence of T7 DNA in at least one of three ways: by deletion mapping (Studier *et al.*, 1979; Studier, 1981), by recombination of T7 mutants with cloned fragments of wild-type T7 DNA (Studier & Rosenberg, 1981; unpublished results), and by direct determination of the nucleotide sequence at the site of the mutation (Dunn *et al.*, 1978; Dunn & Studier, 1981; this paper).

The coding sequences for the 38 known genes do not completely fill the nucleotide sequence of T7 DNA. However, the gaps between known genes are occupied by potential genes, each of which has a recognizably good ribosome-binding and initiation site for protein synthesis (Tables 10 to 12) followed by an open reading frame. Twelve such potential genes have been identified, five in the first 30% of T7 DNA (Dunn & Studier, 1981) and seven in the remaining DNA (Tables 1 to 3 and Fig. 2). The coding sequences of these 50 known and potential genes are close-packed but essentially not overlapping in the nucleotide sequence, an arrangement that strongly suggests all 50 of these genes are expressed during

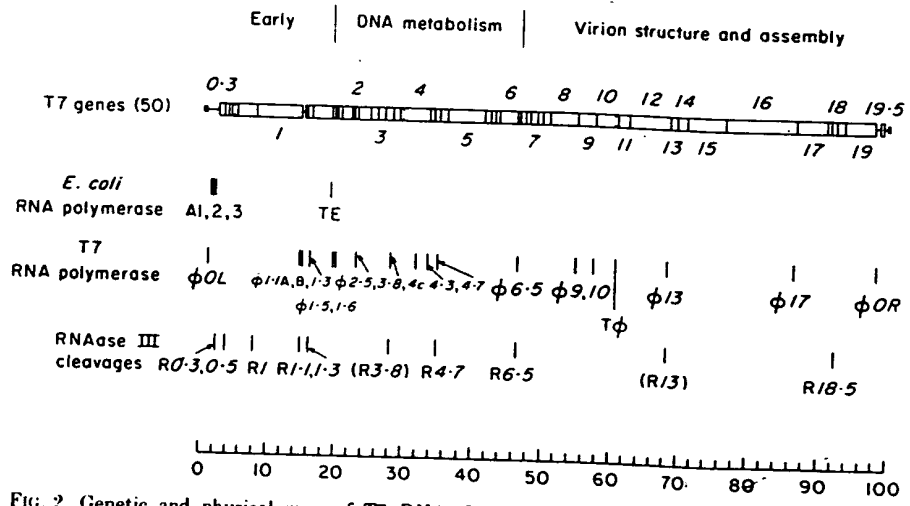


FIG. 2. Genetic and physical map of T7 DNA. The precise positions of the genes and genetic elements in T7 DNA are given in Tables 1 to 3. Coding sequences of the 50 close-packed T7 genes are represented by the open boxes, the terminal repetition by the smaller, filled boxes. The first and last genes, and those with integral numbers are indicated, as are the positions of promoters, transcription termination sites and RNAase III cleavage sites. The scale below represents map units.

infection. The proteins not yet identified are small (29 to 143 amino acids), so it is not surprising that they have escaped detection.

Genes whose coding sequences overlap in different reading frames have been found in bacteriophage ϕX174 (Barrell *et al.*, 1976; Sanger *et al.*, 1978) and might also be expected to occur in T7. We have searched for potential genes whose coding sequences would overlap one or more of the 50 close-packed genes (see Synthesis of T7 Proteins, section 7(b)). Five potential overlapping genes have been identified, which would specify proteins of 40 to 112 amino acids, and preliminary genetic evidence (unpublished results) suggests that one of them is expressed. For convenience, we refer to genes in the above set of 50 as "close-packed", and genes whose coding sequences overlap one of the close-packed genes in a different reading frame as "overlapping".

T7 genes are numbered in order from left to right, according to their position on the genetic map: the 19 genes on the original genetic map have integral numbers (Studier, 1969), and genes added subsequently have decimal numbers. Potential overlapping genes are given numbers according to the relative positions of their left ends. The first gene at the left end of T7 DNA is gene *0·3* and the last gene at the right end is gene *19·5*. (A gene *20* was proposed by Pao & Speyer (1975), but our unpublished results on a mutant kindly provided by Dr Speyer indicate that the proposed gene *20* mutation actually lies in gene *5·7*.)

The 55 known and potential T7 genes we have identified in the nucleotide sequence are listed in order in Table 4, together with the predicted sizes and, where known, functions of the proteins specified. An example of the pattern of protein synthesis during T7 infection is given in Figure 3. T7 genes are expressed co-ordinately in three groups (Studier, 1972): the early, or class I genes are transcribed by *Escherichia coli* RNA polymerase, and include functions to overcome host restriction and to convert the metabolism of the host cell to the production of T7 proteins; the class II genes are the next to be expressed, and include functions involved in DNA metabolism; the class III genes are the last to be expressed, and include genes for proteins of the phage particle and functions involved in maturation and packaging of the DNA. The boundary between class I and II genes is the transcription termination site for *E. coli* RNA polymerase between genes *1·3* and *1·4*; the boundary between class II and class III genes is the ϕ*6·5* promoter for T7 RNA polymerase, located between genes *6·3* and *6·5*.

The sizes predicted for the T7 proteins are generally in good agreement with the sizes estimated from gel electrophoresis in the presence of sodium dodecyl sulfate (Studier, 1972,1973*b*,1981; Studier *et al.*, 1979; see Fig. 3). Most T7 genes appear to specify a single protein, but at least four genes are known or thought to specify two distinct proteins, referred to by the gene number followed by *A* or *B*. These double proteins are discussed in section 7, Synthesis of T7 Proteins.

We proposed a gene *0·65*, based on the presence of a substantial open reading frame but a somewhat unusual ribosome-binding and protein initiation site (Dunn & Studier, 1981). With the discovery that genes *5·5* and *10* apparently produce double proteins by frameshifting during translation (see Synthesis of T7 Proteins section 7(c)), it seems possible that the open reading frame previously assigned to gene *0·65* may in fact be translated primarily by frameshifting during synthesis of

TABLE 4

Known and potential proteins specified by T7 DNA

	Gene [a]	Amino acids [b]	M_r [c]	Function [d]
Class I	*0·3*	116	13.678	Inactivates host restriction
	0·4	50	5621	
	0·5	47	4744	
	0·6A	53	6201	
	0·6B	111	(13.250)	
	0·7	359	41.124	Protein kinase
	1	883	98.092	T7 RNA polymerase
	1·1	42	5180	
	1·2	84	10.059	Replication
	1·3	359	41.133	DNA ligase
Class II	*1·4*	51	5446	
	1·5	29	3174	
	1·6	86	9946	
	1·7	195	22.053	
	1·8	48	5781	
	2	63	7043	Inactivates host RNA polymerase
	2·5	231	25.562	Single-stranded DNA-binding protein
	2·8	139	15.617	
	3	148	17.040	Endonuclease
	3·5	150	16.806	Amidase (lysozyme)
	3·8	121	14.329	
	4A	566	62.656	Primase
	4B	503	55.743	Primase
	(4·1)	39	4265	
	(4·2)	112	12.653	
	4·3	70	7927	
	4·5	88	9960	
	4·7	135	15.208	
	5	704	79.692	DNA polymerase
	5·3	118	13.067	
	5·5	98	11.075	
	5·7	68	7280	Permits growth on λ lysogens
	6	347	39.995	Exonuclease
	6·3	37	4088	
Class III	*6·5*	84	9474	
	6·7	87	9207	
	7	132	15.303	Host range
	7·3	98	9937	Host range
	7·7	130	14.737	
	8	535	58.989	Head–tail protein
	9	306	33.766	Head assembly protein
	10A	344	36.414	Major head protein
	10B	397	(41.800)	Minor head protein
	11	196	22.289	Tail protein
	12	793	89.265	Tail protein
	13	138	15.852	Internal virion protein
	14	195	20.836	Internal virion protein
	15	746	84.210	Internal virion protein
	16	1318	143.840	Internal virion protein
	17	552	61.441	Tail fiber protein
	17·5	67	7391	
	18	89	10.145	DNA maturation

TABLE 4 (*continued*)

Gene[a]	Amino acids[b]	M_r[c]	Function[d]
18·5	143	16,243	
(*18·7*)	82	9195	
19	585	66,130	DNA maturation
(*19·2*)	84	9264	
(*19·3*)	56	6429	
19·5	49	5434	

[a] The gene numbers for the 5 potential overlapping genes are given in parentheses.
[b] The predicted number of amino acids is given, assuming the retention or loss of the initiating methionine residue that is indicated in Tables 1 to 3.
[c] Molecular weights are calculated as described by Dayhoff (1972). The molecular weights predicted for the gene *0·6B* and *10B* proteins are approximate, since the precise point of the predicted translational frameshift has not been identified.
[d] Functions of the T7 proteins are summarized in more detail by Studier (1972,1975*b*). Studier *et al.* (1979). Dunn & Studier (1981), and Saito & Richardson (1981). Gene *5·7* apparently corresponds to the gene described by Pao & Speyer (1975) that permits growth on λ lysogens.

the gene *0·6* protein. This would be consistent with all of the information we have, but the exact nature of the proteins produced from this region remains to be determined. Gene *0·65* is no longer listed as a separate gene.

(b) *Mutations*

The locations in the nucleotide sequence of point mutations in several different genes of the first 30% of T7 DNA have been reported (Dunn *et al.*, 1978; Dunn & Studier, 1981). Several additional point mutations have been located in the remaining sequence. The locations of the mutations and the predicted effects on the proteins are given in Table 5 for all of the mutations so far identified in the nucleotide sequence. All of the amber mutations are in the reading frame predicted by the sequence. The amber mutations in genes *4*, *5* and *6* were chosen for analysis because they were known to lie near the COOH-terminal end of these relatively long proteins and would demonstrate that the predicted reading frame is correct.

The two known mutations in gene *5·5* (Studier, 1981) have both been located in the sequence. As expected, the B64 mutation is an amber mutation, which defines gene *5·5*, but the mutant strain analyzed also contains a missense mutation in gene *5·3*. The B31 mutation, which causes a marked decrease in mobility of the gene *5·5* protein in sodium dodecyl sulfate/polyacrylamide gel electrophoresis (Studier, 1981), changes an arginine residue to glutamine, the only change predicted for the entire gene *5·5* protein. Apparently, a single amino acid change can produce a substantial change in mobility in sodium dodecyl sulfate/polyacrylamide gel electrophoresis, at least for a protein of this size ($M_r = 11{,}075$). (This B31 mutant strain also has a missense mutation in the gene *5·3* protein, and another conceivable explanation for the change in mobility is that the gene *5·3* protein normally processes the gene *5·5* protein.)

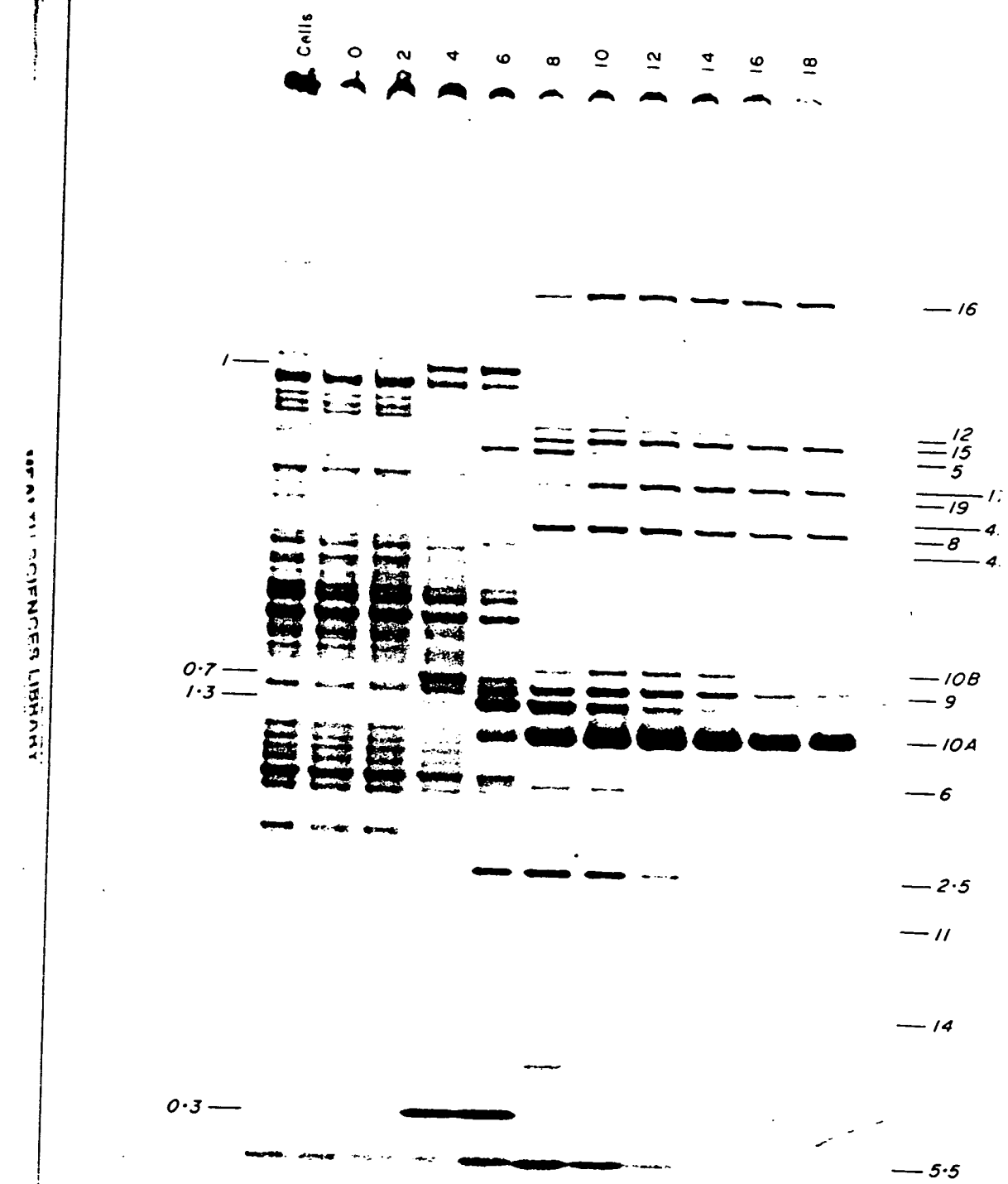

Cells
0
2
4
6
8
10
12
14
16
18
1
0·7
1·3
0·3
16
12
15
5
19
8
10B
9
10A
6
2·5
11
14
5·5

FIG. 3.

TABLE 5

Locations of T7 point mutations in the nucleotide sequence

Mutation	Nucleotide	Change	Amino acid change
0·3-CR35b	914	G to A	
0·3-CR17	926	ATG to ACG	fMet to Thr
0·3-CR17-1	926	ATG to ACG	fMet to Thr
	997	TAT to GAT	Tyr24 to Asp
0·3-CR3b	1189	CAA to TAA	Gln88 to ochre
0·3-CR10b	1168	CTG to TTG	Leu81 unchanged
	1175	GCG to GTG	Ala83 to Val
	1208	TGG to TAG	Trp94 to amber
2-64	8947	TGG to TAG	Trp16 to amber
2-139	9031	TGG to TAG	Trp44 to amber
3-285	10.332	CAG to TAG	Gln25 to amber
3-29	10.506	CAG to TAG	Gln83 to amber
3·5-Lys13a	10.808	CAG to TAG	Gln34 to amber
4-205	13.186	TGG to TAG	Trp541 (478 in *4B*) to amber
5-198	16.351	CAG to TAG	Gln667 to amber
5·5-B64b	16.681	GTT to ATT	Val67 of *5·3* to Ile
	16.953	CAG to TAG	Gln34 to amber
5·5-B31a	16.541	ACC to ATC	Thr20 of *5·3* to Ile
	17.137	CGG to CAG	Arg95 to Gln
6-147	18.328	CAG to TAG	Gln323 to amber
7-36	19.181	TGG to TAG	Trp17 to amber
			creates new *Acc*I site
7-213	19.478	TCT to TTT	Ser116 to Phe
7·3-49	19.538	GGT to GAT	Gly1 to Asp, with
			retention of initiating Met
7·3-405	19.666	CAG to TAG	Gln44 to amber

The mutations have been described by Studier (1969,1975*a*,1981). Silberstein *et al.* (1975) and Dunn *et al.* (1978). The changes in nucleotide sequence for the mutations lying to the left of nucleotide 12,100 were reported by Dunn *et al.* (1978) and Dunn & Studier (1981).

Unexpectedly, the mutations originally assigned to a single gene, gene *7* (Studier, 1969), were found to lie in two adjacent genes, now designated genes *7* and *7·3*. The mutants originally assigned to gene *7* were the only ones in the set defining the 19 genes of the original map for which complementation tests were not possible. Therefore, five mutations that were found to lie between genes *6* and *8* were placed in a single gene simply because they lay in the same region of the map (Studier, 1969). It was subsequently observed, however, that the gene *7* mutants did not all have the same plating behavior on all hosts (Studier, 1975*a*).

FIG. 3. Time course of protein synthesis during T7 infection. A culture of *E. coli* C growing in minimal medium at 30°C was infected with wild-type T7 at a multiplicity of about 15 infective phage particles/cell: 50-μl samples were pulsed for 2 min with 0·5 μCi of [^{35}S]methionine immediately before and at 2 min intervals after infection: the cells were collected by centrifugation, lysed in buffer containing sodium dodecyl sulfate, and subjected to electrophoresis in a sodium dodecyl sulfate-containing Tris/glycine/chloride discontinuous buffer system on a 10% to 20% polyacrylamide gradient gel having a 5% stacking gel, followed by autoradiography, essentially as described by Studier (1973*b*). The origin of electrophoresis is at the top of the patterns. The time at the beginning of each 2 min pulse is given above each lane: lysis of the culture would normally begin about 25 min after infection under these conditions. The gene numbers of prominent T7 proteins are indicated to the side of the patterns, as identified by Studier (1972,1981).

DNA from four of the five original gene *7* mutants has been sequenced in the gene *7* region. Only one change was detected in each mutant: two of the mutations affect gene *7* and two affect gene *7·3*: one amber and one missense mutation was located in each gene. Mutations in gene *7* and *7·3* appear to affect the host range of T7 (Studier; 1975*a*), and the apparent amber phenotype originally observed may reflect a host range property rather than the presence of the amber suppressor.

For the 50 close-packed genes, the predicted proteins that have not yet been identified genetically or biochemically are those specified by genes *1·4*, *1·5*, *1·6*, *1·8*, *3·8*, *5·3*, *6·3*, *6·5*, *6·7*, *7·7*, *18·5* and *19·5*. Deletion mutants that affect the coding sequences of seven of these genes have been isolated (Studier *et al.*, 1979; and unpublished results); the only close-packed genes not yet known to be affected by any available mutation are *6·3*, *6·5*, *6·7*, *18·5* and *19·5*. Proteins whose coding sequences can be deleted can be looked for by comparing the electrophoretic patterns of proteins produced from wild-type or mutant DNA, and preliminary results have tentatively identified two of the predicted proteins. It might also be possible to identify the predicted gene products by use of antibodies raised against synthetic peptides that have amino acid sequences predicted to be in individual proteins, as described by Walter *et al.* (1980) and Sutcliffe *et al.* (1980).

(c) *Packing*

The coding sequences of the 50 close-packed T7 genes occupy 91·9% of the nucleotide sequence (Tables 1 to 3 and Fig. 2), and where any sizable non-coding sequences occur, recognizable genetic signals are almost always found. The longest non-coding stretches occur at the two ends of the DNA; if these are excluded, 95·0% of the internal DNA is coding sequence.

Taking the coding sequence for a protein to extend from the first nucleotide of the initiation codon to the last nucleotide of the termination codon, there are 12 cases among the close-packed genes where the coding sequences of adjacent genes overlap. Seven of these are one base-pair overlaps of the termination codon for one protein with the initiation codon of the next. The other overlaps are 26 base-pairs (between gene *3·8* and the first of the two protein start sites in gene *4*), 20 base-pairs (genes *2·8* and *3*), 14 base-pairs (genes *6* and *6·3*), eight base-pairs (genes *1·7* and *1·8*), and four base-pairs (genes *0·5* and *0·6*). The ribosome-binding and protein initiation sites for 19 of the close-packed T7 proteins lie partly or wholly within the coding sequence for the preceding protein.

Where adjacent coding sequences do not overlap (37 cases), the gap between coding sequences ranges from 0 to 258 base-pairs (average 52). Nineteen of these gaps, 25 to 258 base-pairs long (average 87), contain promoters for T7 RNA polymerase, transcription termination sites, RNAase III cleavage sites, and origins of replication; and six of these 19 gaps contain both a promoter for T7 RNA polymerase and an RNAase III cleavage site (see Fig. 6). These signals are discussed in the following sections. Seventeen of the gaps, 1 to 58 base-pairs long (average 16), contain none of the above genetic signals, nor any others we could recognize, except for ribosome-binding and protein initiation sites.

T7 is another example of the general finding that viruses make very efficient use of the nucleotide sequence available to them. This efficiency is presumably the result of evolutionary packing of the maximum amount of useful information into a DNA whose size is limited by a virion of fixed size.

4. Transcription of T7 DNA

Work from a number of laboratories has produced a comprehensive picture of how T7 DNA is transcribed during infection (for a review, see Dunn & Studier, 1981; McAllister *et al.*, 1981). T7 DNA is transcribed entirely from left to right, first by *E. coli* RNA polymerase, which transcribes the early region, and then by newly-made T7 RNA polymerase, which transcribes a portion of the early region and the entire late region. All of the major promoters and transcription termination sites for both polymerases have now been identified in the nucleotide sequence: their positions in the sequence are given in Tables 1 to 3, and the total transcription pattern is represented in Figure 4.

(a) *Promoters for* E. coli *RNA polymerase*

Three major early promoters for *E. coli* RNA polymerase, A1, A2 and A3, lie in the non-coding region near the left end of T7 DNA (Dunn & Studier, 1973; Minkley & Pribnow, 1973). In addition, several minor promoters, including the leftward promoter AO (also called the D promoter), have been identified by *in vitro* transcription studies (Minkley & Pribnow, 1973; Delius *et al.*, 1973; Stahl

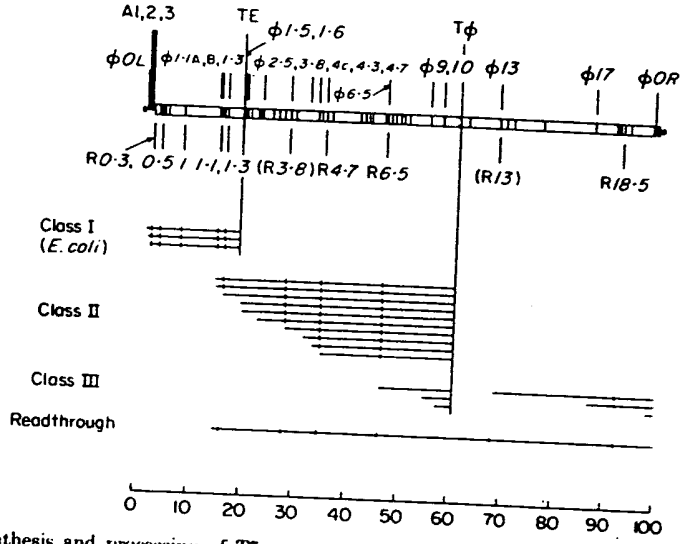


FIG. 4. Synthesis and processing of T7 messenger RNAs. The T7 genes are represented by open boxes; the positions of transcription signals are given above the genes, the RNAase III cleavage sites below. The primary transcript from each promoter is represented by a horizontal line, and sites of RNAase III cleavage by the short vertical lines. Apparently not all RNAs are cut at the R3·8 and R13 RNAase III cleavage sites, as indicated by parentheses. RNAs produced by readthrough of Tφ are also represented. The scale below is in map units.

& Chamberlin, 1977). What role, if any, these minor promoters have in a normal infection has yet to be determined. The start sites for the three major promoters and for the minor A0 and C promoters have been determined directly (Pribnow, 1975; Siebenlist, 1979; McConnell, 1979; Zaychikov & Pletnev, 1980), and the location and start site of the minor B promoter has been inferred (Dunn & Studier, 1981) by matching to sequences of known promoters (Siebenlist *et al.*, 1980). The sequences of these promoters, together with the additional minor promoters discussed below, are gathered in Table 6.

Knowing the total nucleotide sequence of T7 DNA makes it possible to interpret the experiments of Allet *et al.* (1974), in which binding of *E. coli* RNA polymerase to T7 DNA was found to block certain *Hpa*I and *Hin*dII cleavages. Analysis of the restriction patterns in their Figure 3 indicates that cleavages at positions 0·64, 1·79, 46·59, 63·16 and 92·40 were selectively inhibited, the first two strongly, the latter three weakly. The first two sites lie in the A0 and A3 promoters, and the last site can be assigned to the E promoter (Delius *et al.*, 1973; Stahl & Chamberlin, 1977). Examination of the nucleotide sequence around the cleavage site at position 92·40 (the E promoter) identifies a potential rightward promoter, located so that binding of *E. coli* RNA polymerase would be expected to block cleavage. No potential rightward promoters were found near the cleavage sites at 46·59 and 63·16, but potential leftward promoters were identified (Table 6).

Experiments to clone fragments of T7 DNA into plasmids suggest indirectly that several different minor promoters, including the B and C promoters, can be utilized in plasmids (Studier & Rosenberg, 1981; and unpublished results). Promoter activity is inferred because fragments of T7 DNA in which one of these predicted promoters is directed toward a T7 gene cannot be cloned in the *Bam*HI site of plasmid pBR322, whereas fragments that contain the gene but not the promoter can be cloned in the silent but not the expressed orientation, relative to the tetracycline promoter of pBR322. Presumably, the minor promoter can transcribe the T7 gene in the plasmid at a level sufficient to be lethal to the plasmid-containing cell. Examination of the nucleotide sequence in the regions where such promoters would be expected to lie has identified potential promoters near positions 14·26, 23·04, 27·19, 30·50, 41·96 and 86·42 (Table 6). At least one additional promoter should lie between positions 48·8 and 56·8, and three likely candidates have been identified (not shown).

The nucleotide sequences of the 15 known and inferred promoters in Table 6, when compared to the consensus sequence derived from a set of 54 promoters by Siebenlist *et al.* (1980), generally have a good match to the T-T-G-A-C sequence in the −35 region, to the A residue nine nucleotides ahead of this, and to positions 1, 2 and 6 of the T-A-T-A-A-T in the −10 region. There is much less homology to positions 3 to 5 of the T-A-T-A-A-T, as also tends to be the case in the promoters analyzed by Siebenlist *et al.* (1980). The degree of matching of the strong A1, A2 and A3 promoters to the consensus sequence is not strikingly better than that of the minor promoters, suggesting that degree of matching is not a particularly good indicator of relative promoter strength. One feature that does appear to correlate with promoter strength is a string of A residues in the −45 region: A1

TABLE 6

Sequences in T7 DNA that may act as promoters for E. coli *RNA polymerase*

Promoter[a]		Position[b]		−40	−30	−20	−10	1	10
				A	TTGAC		TATAAT		
A0	L	231	0·58	TTTAAGATAGGCGTTGACTTGATGG		GTCTTTAGGTGTAGGCTTTAGGTGTTGGCTTTAGGATG			
A1	R	490	1·23	ATCAAAAAGAGTATTGACTTAAAGT		CTAACCTATAGGATACTTACAGCCATCGAGAGGGACAC			
A2	R	618	1·55	CGAAAAACAGGTATTGACAACATGAA		GTAACATGCAGTAAGATACAAATCGCTAGGTAACACTA			
A3	R	742	1·86	TGAAACAAAACGGTTGACAACATGA		AGTAAACACGGTACGATGTACCACATGAAACGACAGTG			
B	R	1506	3·77	TTATGATTATCACTTTACTTATGAG		GGAGTAATGTATATGCTTACTATCGGTCTACTCACCGC			
C	R	3106	7·78	ATTGATAAGCAACTTGACGCAATGT		TAATGGGCTGATAGTCTTATCTTACAGGTCATCTGCGG			
	R	5695	14·26	CGCGAAACTATGGTTGACACATAT		GAGTCTTGTGATGTACTGGCTGATTTCTACGACCAGTT			
	R	9202	23·04	GGCTAAGAAGATTTTCACCTCTGCGC		TGGGTACCGCTGAACCTTACGCTTACATCGCCAAGCCG			
	R	10,857	27·19	ACAAAGAGCAGGGTTGGCTCGATGTG		GGATACCACTTTATCATCAAGCGAGACGGTACTGTGGA			
	R	12,180	30·50	AACTACGAATACTTTGACCAGTTCGA		ACAGATTATCTTAATGTTCGATATGGACGAAGCAGGGC			
	R	16,758	41·96	AAGGATGGAGACTTTGACATTCTTGCGGTTGTGGTTGACGAAGATGTGCTTATTTTCACATGGGAC					
	L	18,583	46·53	TGGTTAATAGGTGTTAACATAATA		AAGATTCCTCTTAGAGTTAAATCTTAAAGTAATAATCG			
	L	25,198	63·10	AACAACGTTACGGTTAACGATGAAC		GTATAGTCTGCTACAGTAACCATTCGCAGGTCGTTACG			
	R	34,511	86·42	ACCGAACGACCCATTGACTCAACAGC		TTGTGTTGAAGATTTATGAGGCGAACGGTGTTAACTTG			
E	R	36,920	92·45	TGAAATCTGAAAGTTAACGGGAGG		CATTATGCTAGAATTTTTACGTAAGCTAATCCCTTGGG			

The evidence that these sequences have promoter activity, or may have it, is summarized in the text. The nucleotide sequences are for the DNA strand that has the same sequence as the RNA produced from each promoter. The consensus sequence given above the T7 sequences is from known promoters for *E. coli* RNA polymerase (Siebenlist *et al.*, 1980); the equivalent sequences and the known start sites for the promoters in T7 DNA are underlined.

[a] Promoter designations are summarized in the text. L indicates that transcription from the promoter is directed leftward, R that it is directed rightward in the T7 DNA molecule.

[b] The position of the constant T corresponding to position −8 in the consensus sequence is given, since the position of the first nucleotide of the RNA chain is not known for most of these promoters. The sequence hyphens have been omitted for clarity.

and A2 have five A residues in a row, and A3 has four in a row and seven of eight in this region. Several of the minor promoters are comparably rich in A·T base-pairs in this region, but none has the concentration of A residues in the sense (RNA) strand that the major promoters have. Besides the A1, A2 and A3 promoters, perhaps ten of the promoters listed by Siebenlist *et al.* (1980) have strings of A residues in the −45 region. These promoters are generally in phage DNAs or direct the transcription of ribosomal genes, and most of them are known or are likely to be strong promoters. Bound *E. coli* RNA polymerase apparently extends to the −45 region of the promoter (Siebenlist *et al.*, 1980), and perhaps a concentration of A residues in the sense strand in this region increases promoter strength.

(b) *Promoters for T7 RNA polymerase*

Promoters for T7 RNA polymerase have a highly conserved sequence of 23 continuous base-pairs that includes the RNA start site (Oakley & Coleman, 1977; Oakley *et al.*, 1979; Rosa, 1979,1981*b*; Panayotatos & Wells, 1979; Boothroyd & Hayward, 1979; Dunn & Studier, 1981; Carter & McAllister, 1981; this work). Computer search of the total nucleotide sequence of T7 DNA in both directions finds 17 such promoters, all oriented for transcription rightward (Table 7). Individual promoters are identified by a ϕ followed by the number of the gene first transcribed from the promoter, except for ϕOL and ϕOR, which are thought to be parts of replication origins located near the left and right ends of the DNA (Dunn & Studier, 1981; Studier & Rosenberg, 1981). Assays of promoter activity *in vivo* and *in vitro* have shown that all 17 of these promoters can be utilized by T7 RNA polymerase, and have provided no convincing evidence for any other promoters (Golomb & Chamberlin, 1974; Kassavetis & Chamberlin, 1979; Carter *et al.*, 1981; McAllister *et al.*, 1981; Osterman & Coleman, 1981). Thus, these 17 promoters apparently constitute the complete set of promoters for T7 RNA polymerase in T7 DNA.

The promoters for T7 RNA polymerase in T7 DNA have been classified into three groups, referred to as class II, class III, or replication promoters, based on their location, utilization and nucleotide sequence (Carter *et al.*, 1981; Dunn & Studier, 1981; Studier & Rosenberg, 1981; McAllister *et al.*, 1981).

The five class III promoters direct the transcription of class III genes, and appear to be the strongest promoters *in vitro* and *in vivo* (Golomb & Chamberlin, 1974; Niles & Condit, 1975; McAllister & McCarron, 1977; McAllister & Wu, 1978; Kassavetis & Chamberlin, 1979; McAllister & Carter, 1980; McAllister *et al.*, 1981). They all have exactly the same sequence of 23 base-pairs, located at positions −17 to +6 relative to the start site for the RNA.

The ten class II promoters direct the transcription of class II genes. They appear to be weaker than the class III promoters, and differ from the conserved class III promoter sequence in from two to seven positions. Three of the class II promoters, $\phi 1{\cdot}1A$, $\phi 1{\cdot}1B$ and $\phi 1{\cdot}3$, actually lie within the early (class I) region, but they direct transcription into the class II region; and transcription from all of the class II promoters continues into the class III region (see Fig. 4). The $\phi 1{\cdot}1A$ and B promoters are part of the primary origin of replication of T7 DNA (Saito *et al.*,

TABLE 7

Promoters for T7 RNA polymerase

Promoter	RNA start site Nucleotide	RNA start site T7 units	Nucleotide
			−10 1
Conserved sequence			TAATACGACTCACTATGGGAGA
Replication promoter			−20 −10 1
φOL	405	1·01	TTGTCTTTAT TAATACAACTCACTATAAGGAGA GA
Class II promoters			−20 −10 1
φ1·1A	5848	14·64	AACGCCAAAT CAATACGACTCACTATAGAGGGA CA
φ1·1B	5923	14·83	TTCTTCCGGT TAATACGACTCACTATAGGAGGA CC
φ1·3	6409	16·05	GGACTGGAAG TAATACGACTCAGTATAGGGACA AT
φ1·5	7778	19·48	AGTTAACTGG TAATACGACTCACTAAAGGAGGT AC
φ1·6	7895	19·77	TGGTCACGCT TAATACGACTCACTAAAGGAGAC AC
φ2·5	9107	22·80	AGCACCGAAG TAATACGACTCACTATTAGGGAA GA
φ3·8	11.180	27·99	CGTGGATAAT TAATTGAACTCACTAAAGGGAGA CC
φ4c	12.671	31·73	CCGACTGAGA CAATCCGACTCACTAAAGAGAGA GA
φ4·3	13.341	33·41	AGTCCCATTC TAATACGACTCACTAAAGGAGAC AC
φ4·7	13.915	34·84	TTCATGAATA CTATTCGACTCACTATAGGAGAT AT
Class III promoters			−20 −10 1
φ6·5	18.544	46·43	GTCCCTAAAT TAATACGACTCACTATAGGGAGA TA
φ9	21.864	54·75	GCCGGGAATT TAATACGACTCACTATAGGGAGA CC
φ10	22.903	57·35	ACTTCGAAAT TAATACGACTCACTATAGGGAGA CC
φ13	27.273	68·29	GGCTCGAAAT TAATACGACTCACTATAGGGAGA AC
φ17	34.565	86·55	GCGTAGGAAA TAATACGACTCACTATAGGGAGA GG
Replication promoter			−20 −10 1
φOR	39.228	98·23	CACGATAAAT TAATACGACTCACTATAGGGAGA GG

The 17 promoters are grouped as described in the text. The nucleotide sequences are from the *l* strand of T7 DNA. The RNA start site is assumed to be at the equivalent position in each promoter. The 23-base conserved sequence is set off from adjacent sequences by a space, and nucleotides that deviate from the class III promoter sequence are identified by asterisks. In the conserved sequence at the top of the Table, nucleotides that are unchanged in all 17 promoters are indicated by a line above the letter, those that are found in 16 of the 17 promoters are indicated by a dot, and the first nucleotide of the RNA chain is underlined. The sequence hyphens have been omitted for clarity.

1980: and our published results), and could therefore be considered to be replication promoters as well as class II promoters.

The *φOR* promoter is considered to be a replication promoter because it lies within a fairly long non-coding region (258 base-pairs) and is part of a replication origin (unpublished results). However, it might also be considered a class III promoter: it directs transcription of one small gene, gene *19·5*: it completely matches the conserved class III promoter sequence: and it appears to be utilized as efficiently as the class III promoters *in vitro* (Golomb & Chamberlin, 1974). The *φOL* promoter, also thought to be part of a replication origin (Dunn & Studier, 1981), differs from the conserved class III promoter sequence in two positions (Table 7). Although the *φOL* promoter can be utilized in a restriction fragment of T7 DNA (Osterman & Coleman, 1981), and may indeed direct transcription of the entire early region *in vitro* (Scherzinger *et al.*, 1972), there is no evidence that *φOL* directs transcription of the early genes of T7 *in vivo* (McAllister & Wu, 1978: McAllister *et al.*, 1981: our unpublished results).

Uninterrupted homology between pairs of promoters can continue past the conserved sequence of 23 base-pairs: the *φ10* and *φ13* promoters are identical for 30 consecutive base-pairs (−24 to +6): the *φ6·5* and *φOR* promoters match each other for 28 continuous base-pairs (−22 to +6) and match the *φ10* and *φ13* promoters for 27 consecutive base-pairs (−21 to +6): the *φ9* and *φ10* promoters have a continuous match of 26 base-pairs (−18 to +8): and the *φ17* and *φOR* promoters have a match of 25 base-pairs (−17 to +8). The longest perfect match involving class II promoters is between the *φ1·6* and *φ4·3* promoters, 25 continuous base-pairs from −17 to +8.

The 11 promoters that differ from the conserved class III promoter sequence do so mostly in positions +3 to +6: only 50% of these bases match the class III sequence, whereas over 90% of the bases in positions −17 to +2 match the class III sequence. What appears to be conserved in the region past the start site is the presence of purines in the *l* strand: 92% of the bases in positions −1 to +6 of these 11 promoters are A or G, and all 17 promoters have an uninterrupted string of 5 to 12 purines that includes the start site for the RNA. Considering all 17 promoters, nine positions are completely conserved and four more are the same in all but one promoter (Table 7). These 13 most highly conserved positions lie between −16 and −1, suggesting that the precise sequence of this region, together with a polypurine tract in the −1 to +6 region, are important factors in defining a promoter.

Panayotatos & Wells (1979), Rosa (1979), and Osterman & Coleman (1981) have reported that removal of DNA to the left of the *Hpa*II site at position −21 relative to the *φ1·1*B start site, to the left of the *Hpa*II site at position −24 relative to the *φ9* start site, or to the left of the *Taq*I site at position −23 relative to the *φ13* start site, has little effect on promoter activity for purified T7 RNA polymerase. Apparently, promoter activity requires no sequence information, nor any DNA at all to the left of position −21 or so. However, removal of the DNA to the left of position −10, by cutting at the *Hin*fI site found in most promoters, abolishes promoter activity (Rosa, 1979: Oakley *et al.*, 1979: Osterman & Coleman, 1981). Osterman & Coleman (1981) found that replacement of the DNA

to the left of the *HinfI* site by two different sequences from non-promoter fragments could restore promoter activity, suggesting that the precise sequence to the left of position −11 is not critical. However, we have replaced the sequence to the left of −11 with yet another sequence, and find no promoter activity *in vivo* (unpublished preliminary result), so some sequence information may be needed in this region.

It is reasonable to suppose that the differences between the class II promoter sequences and the conserved 23-base class III promoter sequence are responsible for the relative weakness of class II promoters *in vivo* and *in vitro*. However, a further distinction has been pointed out by Rosa (1979,1981*b*) and by McAllister & Carter (1980): the class III promoters have an uninterrupted string of A·T base-pairs extending leftward from position −13 for eight to ten base-pairs, whereas this string is typically interrupted after only three to six base-pairs in the class II promoters. The only exception among the class II promoters is the *ϕ3·8* promoter, which has an uninterrupted string of ten A·T base-pairs. The *ϕ3·8* promoter appears to be a relatively weak promoter, but it continues to be utilized *in vivo* longer than is typical for class II promoters (McAllister *et al.*, 1981). Perhaps the changes within the conserved sequence are responsible for the weakness of the promoter but the string of A·T base-pairs makes it able to compete for T7 RNA polymerase at late times better than other class II promoters. Interestingly, however, the *ϕ3·8* promoter is also the only class II promoter that completely matches the conserved class III sequence at positions −1 to +6. The replication promoters *ϕOL* and *ϕOR* are like the class III promoters in having uninterrupted strings of ten and 11 A·T base-pairs extending leftward from position −13.

Alignment of the 17 promoter sequences (Table 7) shows that all but two have G at the position corresponding to the RNA start sites that have been determined (Rosa, 1979; Oakley *et al.*, 1979). The two exceptions are the *ϕOL* and *ϕ2·5* promoters, which have A at this position. RNA chains transcribed from T7 DNA by T7 RNA polymerase are known to begin almost entirely with GTP (Chamberlin & Ring, 1973), but a small amount of initiation with ATP would be consistent with the data. We have observed incorporation of [γ-^{32}P]ATP into RNA initiated at the *ϕ2·5* promoter (not shown), and therefore we presume that RNAs from all of the promoters start at position +1 in the alignment shown in Table 7. (See note added in proof.)

All but one of the promoters for T7 RNA polymerase in T7 DNA are located in sequences that do not code for any protein. The exception is the *ϕ4c* promoter (so named because it lies within the coding sequence for gene *4*). The smallest non-coding gaps that contain promoters are 25 base-pairs (*ϕ1·6*) and 27 base-pairs (*ϕ1·5*). All of the other non-coding gaps that contain promoters are at least 63 base-pairs long.

(c) *Transcription termination sites*

Transcription of the early region of T7 DNA by *E. coli* RNA polymerase ends at one of two adjacent nucleotides at position 19·00: about two-thirds of the RNA chains end at the C at nucleotide 7588 and one-third at the following G (Dunn &

Studier, 1980). The nucleotides in the RNA preceding the terminal nucleotides can be arranged in a relatively stable stem-and-loop structure that contains eight uninterrupted base-pairs and a four-base loop (Fig. 5), a structure commonly found in transcription termination sites for *E. coli* RNA polymerase (Rosenberg & Court, 1979). Termination at this site is rho-independent and is not completely efficient *in vivo* or *in vitro* (Millette *et al.*, 1970; Studier, 1972; Kiefer *et al.*, 1977). We refer to this transcription termination site as TE (terminator, early). There is evidence that transcription by *E. coli* RNA polymerase can terminate specifically at two additional sites in the late region of T7 DNA: first, near position 28 (Minkley & Pribnow, 1973; Peters & Hayward, 1974), and second, at the transcription termination site for T7 RNA polymerase near position 60·6 (unpublished results). Termination at these sites appears to be less efficient than at TE.

Transcription by T7 RNA polymerase does not terminate at TE, but proceeds through it into the late region (Dunn & Studier, 1980). One strong termination site for T7 RNA polymerase has been identified in T7 DNA, between genes *10* and *11* (Niles & Condit, 1975; McAllister & McCarron, 1977). (A proposed termination site near position 98·5 (Golomb & Chamberlin, 1974) has not been confirmed.) The RNA nucleotides in the non-coding gap between genes *10* and *11* can be arranged in a stem-and-loop structure that contains 14 to 15 uninterrupted base-pairs and a six or eight-base loop (Fig. 5). This stem ends in a string of six consecutive U residues, the only place in the entire T7 DNA molecule where as many as six consecutive T residues are found. The RNA chain ends at the G at nucleotide 24,209 (position 60·62), the first nucleotide past the string of U residues

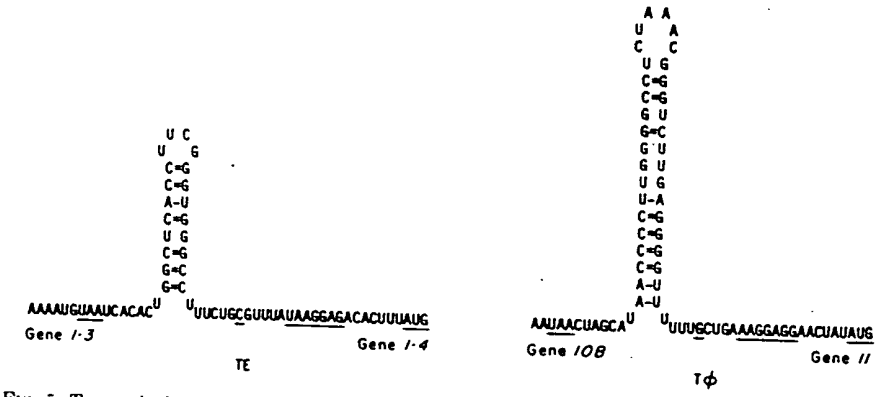


FIG. 5. Transcription termination sites for *E. coli* and T7 RNA polymerases. Potential base-paired structures ahead of the site of termination are indicated. *E. coli* RNA polymerase terminates RNA chains in TE mainly at the underlined C, but about 1/3 of the time at the next G residue (Dunn & Studier, 1980). *E. coli* RNA polymerase also appears to terminate RNA chains at Tφ, but the precise position is not known (unpublished results). T7 RNA polymerase does not terminate RNA chains at TE (Dunn & Studier, 1980), but terminates at the underlined G in Tφ (M. D. Rosa, personal communication). The termination codon for the gene immediately preceding each termination site, and the ribosome-binding sequence and initiation codon for the gene immediately beyond the termination site, are underlined. Sequence hyphens have been omitted but hyphens and double hyphens represent A·U and G·C base-pairs, respectively.

(M. D. Rosa: personal communication). We refer to this transcription termination site for T7 RNA polymerase as Tϕ (terminator. phage).

Termination at Tϕ is not completely efficient *in vitro* (McAllister & McCarron, 1977; Carter *et al.*, 1981: see Fig. 7) or *in vivo* (McAllister *et al.*, 1981). In fact, completely efficient termination at this site would be lethal to T7: there is no promoter for T7 RNA polymerase between the termination site and genes *11* and *12*, which specify structural proteins of the T7 tail, so these genes must be transcribed entirely by readthrough of Tϕ (see Fig. 4).

A termination site just past gene *10* is apparently part of a transcriptional strategy that ensures production of large amounts of the major capsid protein of T7, specified by gene *10*. Transcription from all ten of the class II promoters and from three class III promoters would be expected to cross gene *10* (Fig. 4). Large amounts of gene *10* mRNA are needed, but equivalent amounts of downstream mRNAs might well be deleterious, and a partial termination site behind gene *10* may provide a proper balance.

5. RNAase III Cleavage Sites

Cleavage of T7 RNAs at specific sites by a host enzyme, RNAase III, is a prominent feature of both early and late transcription (Dunn & Studier, 1973,1975). The precise points of cleavage have been determined for the five cleavage sites in the early region (Kramer *et al.*, 1974; Rosenberg *et al.*, 1974; Rosenberg & Kramer, 1977; Robertson *et al.*, 1977; our unpublished results). The RNA around each of these sites can be arranged in a characteristic pattern of base-pairing within which lies the point of cleavage (Fig. 6). In a change from our previous nomenclature for RNAase III cleavage sites (Dunn & Studier, 1981), we now refer to each site by R followed by the number of the first gene to the right of the cleavage site. In this nomenclature, the five early RNAase III cleavage sites are R*0·3*, R*0·5*, R*1*, R*1·1* and R*1·3*.

RNAase III cleavage sites in the T7 late RNAs have been less well mapped, but indirect evidence suggested that sites should lie at least between genes *3·5* and *4*, between genes *6* and *7*, between genes *12* and *13*, and between genes *17* and *19* (Dunn & Studier, 1975,1980; Studier, 1975*b*; Pachl & Young, 1978). Analysis of the pairing potential of nucleotide sequences in non-coding gaps between the late genes, and comparison with the early RNAase III cleavage sites, identifies likely cleavage sites ahead of genes *4·7*, *6·5* and *18·5* (Fig. 6). Other possible cleavage sites might be expected to lie within extensively paired regions ahead of genes *8*, *9*, *10*, *13* and *17*, and perhaps at the transcription termination site between genes *10* and *11*. (We have shown (Dunn & Studier, 1980) that the transcription termination site for *E. coli* RNA polymerase at the end of the early region is not a cleavage site for RNAase III.) In most of these potential cleavage sites, a promoter sequence for T7 RNA polymerase overlaps the sequence thought to be needed for RNAase III cleavage.

To test which potential RNAase III cleavage sites in the late region are actually used, we transcribed plasmid DNAs that contained cloned fragments of T7 DNA, and tested whether the transcripts could be cut by purified RNAase III

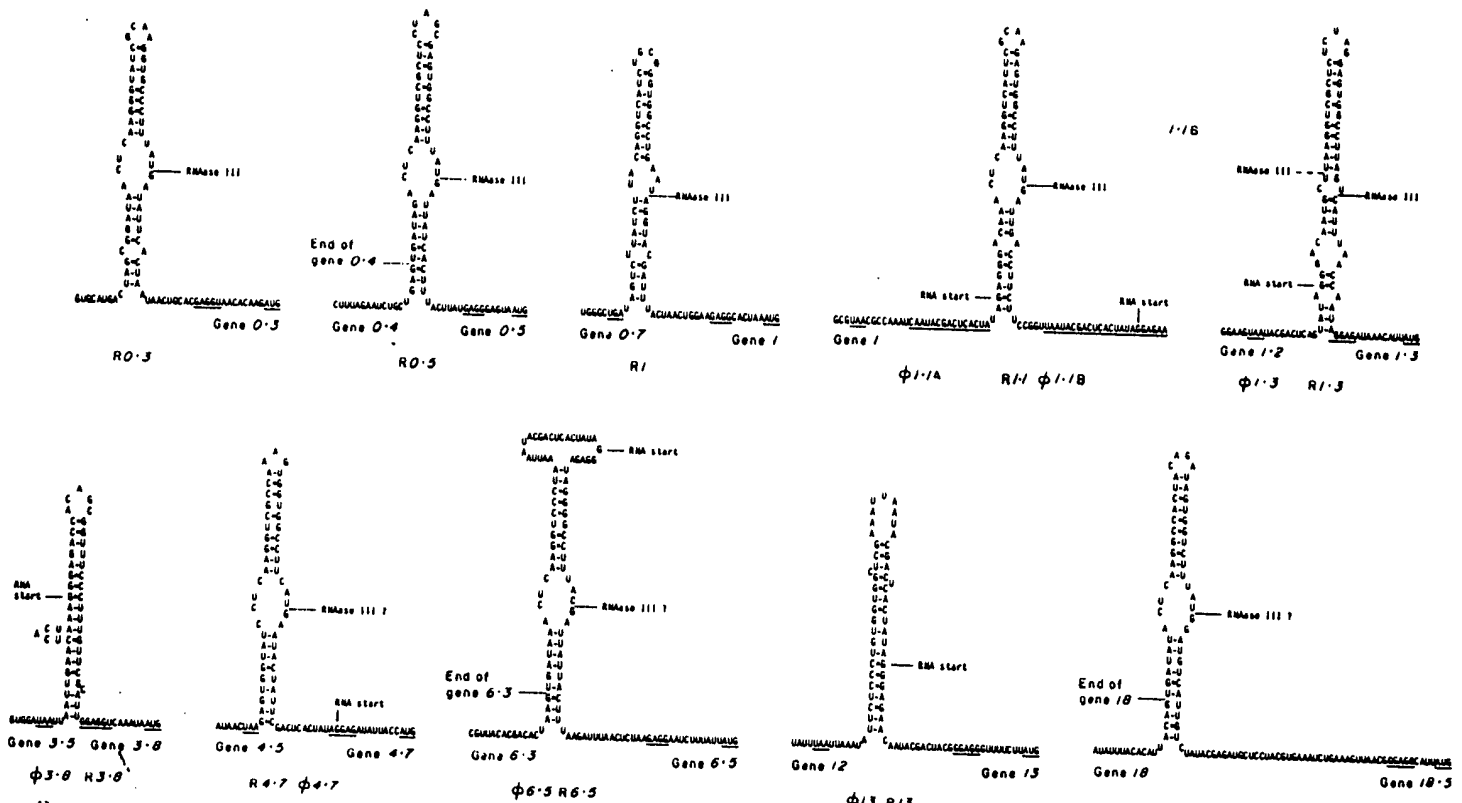


FIG. 6. RNAase III cleavage sites in T7 RNAs. Potential base-paired structures are indicated, as are the locations of the termination codons for the genes immediately preceding, and the ribosome-binding sequences and initiation codons for the genes immediately beyond the cleavage sites. The locations of promoters for T7 RNA polymerase are also shown. The known positions of RNAase III cleavage for the 5 cleavages in T7 early RNA are indicated, as are the predicted positions of cleavage for *R4·7*, *R6·5* and *R18·5*. The position of the secondary cleavage in *RI·3* (Dunn, 1976; Robertson *et al.*, 1977) is indicated by a broken line. Cleavage at (or near) *R3·8* and at *R13* appears to be relatively inefficient, and it is not clear where the cleavage might be expected to occur. Sequence hyphens have been omitted but hyphens and double hyphens represent A–U and G–C base pairs.

(Fig. 7). Each of the plasmids contained a promoter for T7 RNA polymerase oriented so as to direct transcription in the silent (counter-clockwise) direction from the *Bam*HI site of pBR322. a direction in which there are no strong termination sites for T7 RNA polymerase in pBR322 DNA (McAllister *et al.*. 1981). Except for pAR436, which carries the Tφ termination site from T7 DNA. transcription of the plasmids by T7 RNA polymerase produces a heterogeneous

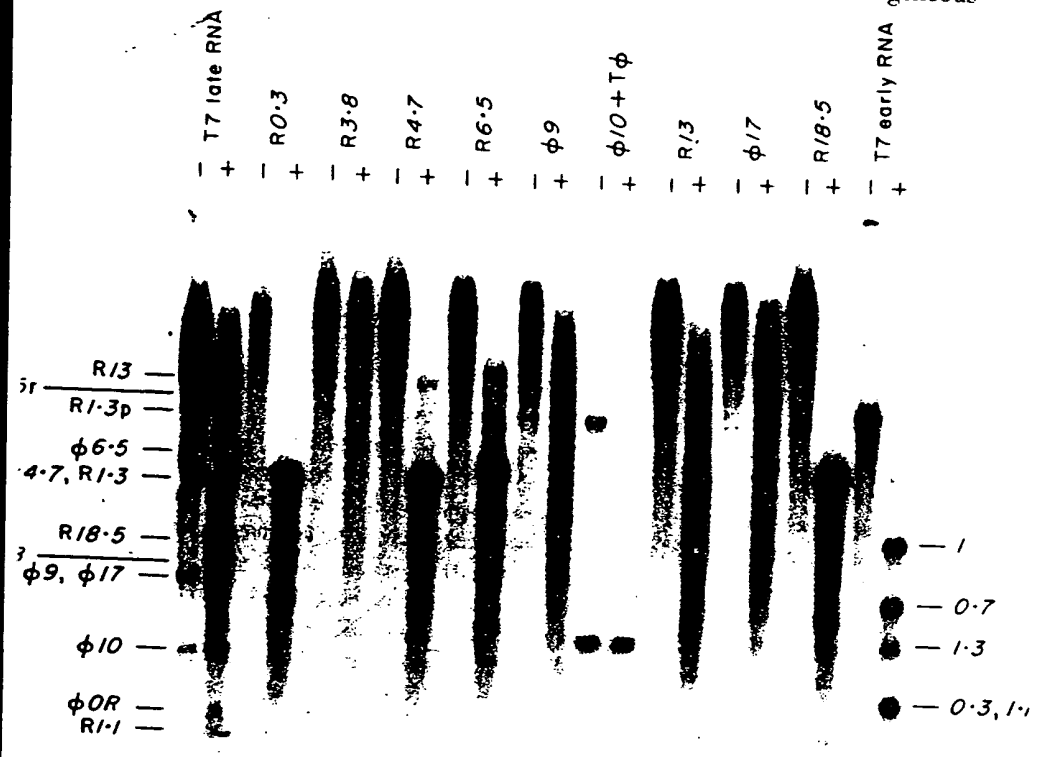


FIG. 7. RNAase III cleavage of RNAs transcribed from recombinant plasmids. T7 DNA or recombinant plasmid DNAs were transcribed by purified T7 RNA polymerase (a gift from C. Fuller & C. Richardson) in the presence of [α-^{32}P]UTP. and a portion was treated with purified RNAase III in a buffer containing 0·2 M-NH_4Cl. essentially as described by Dunn (1976). RNA samples were precipitated with ethanol. heated for 5 min at 65°C in 50 mM-phosphate buffer containing 1·1 M-formaldehyde. and subjected to electrophoresis through a 1% (w/v) agarose slab gel at 2 V/cm in the same solvent. followed by autoradiography. The recombinant plasmids contained fragments of T7 DNA inserted into the *Bam*HI site of pBR322 in the silent (counter-clockwise) orientation (Studier & Rosenberg. 1981). Two of the fragments contained no promoter for T7 RNA polymerase. and these were cloned along with the *φ10* promoter. in the proper orientation. Samples are arranged on the gel in pairs. the second having been treated with RNAase III. The first and last pairs of samples contained RNA transcribed from T7 DNA itself. the first by T7 RNA polymerase. the last by *E. coli* RNA polymerase. Identifications of T7 late and early RNAs. based on relative sizes. are given at the sides of the patterns (see Fig. 4. Tables 8 and 9). The plasmid designations. the locations in T7 DNA of the cloned fragments. and the promoters and cleavage sites within the cloned fragments were: pAR946. 2·49 to 3·45. *φ10* + R*0·3*: pAR410. 26·70 to 28·29. *φ3·8* + R*3·8*: pAR95. 34·39 to 35·45. *φ4·7* + R*4·7*: pAR35. 44·49 to 46·59. *φ6·5* + R*6·5*: pAR213. 49·78 to 56·65. *φ9*: pAR436 57·23 to 60·78. *φ10* + Tφ: pAR665. 67·82 to 68·80. *φ13* + R*13*: pAR651. 85·92 to 87·01. *φ17*: pAR965. 91·41 to 92·40. *φ10* + R*18·5*.

population of RNAs longer than the plasmid DNA. If an RNAase cleavage site is present in the RNA transcribed from the inserted fragment of T7 DNA, the RNAs are cut to a homogeneous length about the size of the plasmid DNA, but if no cleavage site is present, no specific cleavage products are observed (Fig. 7). In the case of pAR436, specific RNAs are produced by termination at Tϕ, but since termination is not completely efficient, specific RNAs that terminate at Tϕ after having transcribed two or three times around the plasmid DNA are also observed. If RNAase III were to cut in either the *ϕ10* region or near Tϕ, a specific RNA somewhat shorter than the RNA produced after one readthrough of Tϕ would be observed.

The results of Figure 7 clearly identify three primary RNAase III cleavage sites in the T7 late RNAs, which can be assigned to the regions ahead of genes *4·7*, *6·5* and *18·5*. In addition, a less efficient cleavage site is observed, which can be assigned to the region ahead of gene *13*. These experiments provided no evidence for RNAase III cleavage sites in the extensively paired regions ahead of genes *3·8*, *9*, *10* or *17*, or at the Tϕ transcription termination site.

Even though no evidence for an RNAase III cleavage site ahead of gene *3·8* was found from transcribing plasmid DNAs, at least three other types of evidence point to a cleavage site in RNA near position 28 in T7 DNA: (1) RNAase III treatment releases a small RNA fragment (approximately 35 nucleotides) from RNA transcribed by *E. coli* RNA polymerase from mutant T7 DNAs that lack the transcription termination site at the end of the early region (Dunn & Studier, 1975). Mapping of the cleavage products suggests that this small RNA arises from near position 28. (2) Appropriate *in vitro* transcripts of full-length or fragmented T7 DNA, when cut with RNAase III, produce specific sizes of RNAs consistent with a cleavage site in this region (results not shown). (3) The effects of T7 deletions on the sizes of the late RNAs produced *in vivo* indicate that a portion but not all of the *in vivo* RNAs are cut in this region (results not shown). Thus, it seems very likely that a relatively weak RNAase III cleavage site lies near position 28.

Examination of potential base-paired structures in all of the non-coding regions between genes *2·5* and *4* (positions 24·5 to 29·0) suggests that the region ahead of gene *3·8* is the most likely to be an RNAase III cleavage site. One possible paired structure for the RNA from this region is shown in Figure 6, but alternative structures are also possible. Why cleavage at or near position 28 was apparent in the transcripts from full-length or fragmented T7 DNA but not in the transcripts from the plasmid DNA is not known: perhaps the cleavage site near position 28 is not ahead of gene *3·8* but is at a nearby site not contained within the plasmid, or perhaps some long-range interaction involving sequences not cloned in the plasmid is needed for cleavage, or perhaps interaction between the RNA produced from the cloned T7 DNA and some sequence in the RNA produced from the plasmid DNA can interfere with cleavage at this site. We tentatively assign a weak RNAase III cleavage site to the region ahead of gene *3·8*, but this will have to be confirmed by further work.

Considering all of the available evidence, it seems that T7 late RNAs contain three efficient RNAase III cleavage sites (R*4·7*, R*6·5* and R*18·5*) and two

relatively inefficient cleavage sites (*R3·8* and *R13*). These, together with the five efficient cleavage sites in the early region (*R0·3*, *R0·5*, *R1*, *R1·1* and *R1·3*), make a total of ten known RNAase III cleavage sites in T7 RNAs. The precise point of cleavage has not yet been determined for any of the late cleavage sites. However, comparison with the cleavage sites in the early region suggests likely points of cleavage for *R4·7*, *R6·5* and *R18·5*, which are indicated in Figure 6.

6. T7 Messenger RNAs

The combined effects of transcription and RNAase III cleavages are predicted to produce a rather large set of late mRNAs (Fig. 4). The relative amounts of individual mRNAs are expected to differ widely, because of differences in promoter strengths and differences in the number of promoters that direct transcription across different regions of T7 DNA, and also because a fraction of RNA chains do not terminate at Tϕ and/or escape RNAase III cleavage at *R3·8* or *R13*. Several pairs of RNAs are predicted to differ slightly only at their 5′ ends, depending on whether the RNA began at a promoter or at an RNAase III cleavage in the same non-coding interval. The positions and sizes of all of the T7 RNAs predicted to be produced from the promoters, cleavage sites and termination sites given in Tables 1 to 3 are listed in Tables 8 and 9: Table 8 lists the RNAs expected in the absence of RNAase III cleavage (conditions frequently used *in vitro*): Table 9 lists the RNAs expected to be produced in the presence of RNAase III cleavage (the normal situation *in vivo*).

In order to refer to any T7 RNA unambiguously, we have adopted a systematic nomenclature based on the promoters, RNAase III cleavage sites, and/or termination sites from which the RNAs were produced (see the legends to Tables 8 and 9). The RNAs that correspond to species II to VI in the nomenclature used by Golomb & Chamberlin (1974) are identified in the Tables. Note that the *ϕ6·5*, *ϕ9*, *ϕ10* and *ϕOR* RNAs are the only T7 RNAs expected to be unaffected by RNAase III cleavage.

T7 RNAs are relatively stable *in vivo* (Summers, 1969,1970: Marrs & Yanofsky, 1971), and the distribution of sizes predicted in Tables 8 and 9 agrees well with the observed distributions produced *in vivo* and *in vitro* (Summers, 1969: Golomb & Chamberlin, 1974; Dunn & Studier, 1975: Pachl & Young, 1978: our unpublished data; see Fig. 7). Late RNAs that appear to be identifiable *in vivo* are identified by asterisks in Table 9. Many of the other late RNAs are presumably made in such small amounts that they would not be readily detectable. Despite the good agreement between the predicted and observed RNA distributions, at least one or two discrepancies may still remain (unpublished work).

Although T7 mRNAs are quite stable, *E. coli* mRNA appears to be degraded as efficiently during T7 infection as in uninfected cells (Marrs & Yanofsky, 1971). This implies that the stability of T7 mRNAs is inherent in the RNA and is not due to an alteration in the ability of the cell to degrade mRNAs during infection. A consistent feature of T7 mRNAs is the potential for a relatively stable base-paired structure at the 3′ end, which would be found in all RNAs that terminate at TE or Tϕ, or which are cut by RNAase III at their 3′ ends (Figs 5 and 6). The only possible exceptions are the *R18·5* and *ϕOR* RNAs, which presumably end at

TABLE 8

Predicted T7 RNAs, unprocessed by RNAase III

	RNA[a]	Position of RNA[b]		Length of RNA		
		Left	Right	Nucleotides	T7 units	M_r (Na^+ salt)
E. coli RNA polymerase						
	A1t	498	7588	7091	17·76	2.442.000
	A2t	626	7588	6963	17·44	2.397.000
	A3t	750	7588	6839	17·12	2.354.000
T7 RNA polymerase						
	*φOL*t	405	24.209	23.805	59·61	8.199.000
	*φ1·1*At	5848	24.209	18.362	45·98	6.325.000
	*φ1·1*Bt	5923	24.209	18.287	45·79	6.299.000
	*φ1·3*t	6409	24.209	17.801	44·57	6.131.000
	*φ1·5*t	7778	24.209	16.432	41·15	5.660.000
	*φ1·6*t	7895	24.209	16.315	40·85	5.620.000
	*φ2·5*t	9107	24.209	15.103	37·82	5.203.000
	*φ3·8*t	11.180	24.209	13.030	32·63	4.488.000
	*φ4*ct	12.671	24.209	11.539	28·89	3.973.000
	*φ4·3*t	13.341	24.209	10.869	27·22	3.743.000
	*φ4·7*t	13.915	24.209	10.295	25·78	3.546.000
IIIa	*φ6·5*	18.544	24.209	5666	14·19	1.952.000
IV	*φ9*	21.864	24.209	2346	5·87	808.400
V	*φ10*	22.903	24.209	1307	3·27	450.000
	*φOL*rt	405	39.936	39.532	98·99	13.610.000
	*φ1·1*Art	5848	39.936	34.089	85·36	11.740.000
	*φ1·1*Brt	5923	39.936	34.014	85·17	11.710.000
	*φ1·3*rt	6409	39.936	33.528	83·95	11.540.000
	*φ1·5*rt	7778	39.936	32.159	80·53	11.070.000
	*φ1·6*rt	7895	39.936	32.042	80·23	11.030.000
	*φ2·5*rt	9107	39.936	30.830	77·20	10.620.000
	*φ3·8*rt	11.180	39.936	28.757	72·01	9.900.000
	*φ4*crt	12.671	39.936	27.266	68·27	9.385.000
	*φ4·3*rt	13.341	39.936	26.596	66·60	9.155.000
	*φ4·7*rt	13.915	39.936	26.022	65·16	8.958.000
	*φ6·5*rt	18.544	39.936	21.393	53·57	7.363.000
	*φ9*rt	21.864	39.936	18.073	45·25	6.220.000
	*φ10*rt	22.903	39.936	17.034	42·65	5.862.000
II	*φ13*t	27.273	39.936	12.664	31·71	4.360.000
IIIb	*φ17*t	34.565	39.936	5372	13·45	1.848.000
VI	*φOR*	39.228	39.936	709	1·78	243.200

[a] RNAs produced in the absence of RNAase III are referred to by the promoter from which they originated: the suffix t identifies RNAs that end at the first appropriate termination site (TE, T*φ*, or the right end of T7 DNA), after having passed through one or more RNAase III cleavage sites: the suffix rt identifies RNAs that were produced by transcription through T*φ* and that end at the right end of the DNA: the promoter designation itself, without a suffix, refers to RNAs that end at T*φ* or at the right end of T7 DNA without having passed through any RNAase III cleavage sites (see Fig. 4). The RNAs that correspond to species II to VI described by Golomb & Chamberlin (1974) are indicated.

designation itself (without any suffix) refers to an RNA that has its right end at the first RNAase III cleavage site or termination site: the suffix r identifies RNAs produced by readthrough of T*φ* and which end at R*13*: the suffix p refers to an RNA that escaped cleavage by RNAase III (presumably only at R*3·8* or R*13*) and which therefore ends at the second RNAase III cleavage site from the left end of the RNA (see Fig. 4). The RNAs that correspond to species IIIa to VI described by Golomb & Chamberlin (1974) are identified. The *φ6·5*, *φ9*, *φ10* and *φOR* RNAs are unaffected by RNAase III cleavage and are therefore identical to the RNAs of the same designation in Table 8.

[b] The nucleotide numbers of the predicted first and last nucleotides in the RNA chain are given.

TABLE 9

Predicted T7 RNAs. after processing by RNAase III

	RNA[a]	Position of RNA[b] Left	Right	Length of RNA Nucleotides	T7 units	M_r (Na^+ salt)
E. coli RNA polymerase						
	A1	498	890	393	0·98	135.700
	A2	626	890	265	0·66	91.400
	A3	750	890	141	0·35	48.580
	0·3	891	1468	578	1·45	198.800
	0·7	1469	3138	1670	4·18	575.600
	1	3139	5887	2749	6·88	945.800
	1·1	5888	6448	561	1·40	193.300
	1·3	6449	7588	1140	2·85	392.600
T7 RNA polymerase						
	φ0L	405	890	486	1·22	167.700
	φ1·1A	5848	5887	40	0·10	13.860
	R*1·1*	5888	6448	561	1·40	193.300
	φ1·1B *	5923	6448	526 *	1·32 *	181.400
	φ1·3	6409	6448	40	0·10	13.850
	R*1·3* *	6449	11.203	4755 *	11·91 *	1.638.000
	φ1·5	7778	11.203	3426	8·58	1.180.000
	φ1·6	7895	11.203	3309	8·29	1.140.000
	φ2·5	9107	11.203	2097	5·25	723.200
	R*1·3*p *	6449	13.892	*7444 *	18·64 *	2.564.000
	*φ1·5*p	7778	13.892	6115	15·31	2.107.000
	*φ1·6*p	7895	13.892	5998	15·02	2.067.000
	*φ2·5*p	9107	13.892	4786	11·98	1.650.000
	φ3·8	11.180	13.892	2713	6·79	934.800
	R*3·8* *	11.204	13.892	2689 *	6·73 *	926.600
	φ4c	12.671	13.892	1222	3·06	420.100
	φ4·3	13.341	13.892	552	1·38	189.900
	R*4·7*	13.893	18.562	4670	11·69	1.608.000
	φ4·7 *	13.915	18.562	4648 *	11·64 *	1.601.000
IIIa	*φ6·5*	18.544	24.209	5666	14·19	1.952.000
	R*6·5* *	18.563	24.209	5647 *	14·14 *	1.945.000
IV	*φ9* *	21.864	24.209	2346 *	5·87 *	808.400
V	*φ10* *	22.903	24.209	1307 *	3·27 *	450.000
	*φ6·5*r	18.544	27.280	8737	21·88	3.006.000
	R*6·5*r *	18.563	27.280	8718 *	21·83 *	3.000.000
	*φ9*r	21.864	27.280	5417	13·56	1.863.000
	*φ10*r	22.903	27.280	4378	10·96	1.505.000
	*φ6·5*rp	18.544	36.855	18.312	45·85	6.304.000
	R*6·5*rp	18.563	36.855	18.293	45·81	6.298.000
	*φ9*rp	21.864	36.855	14.992	37·54	5.161.000
	*φ10*rp	22.903	36.855	13.953	34·94	4.803.000
	φ13	27.273	36.855	9583	24·00	3.301.000
	R*13* *	27.281	36.855	9575 *	23·98 *	3.298.000
	φ17 *	34.565	36.855	2291 *	5·74 *	788.800
	R*18·5* *	36.856	39.936	3081 *	7·71 *	1.059.000
VI	*φOR* *	39.228	39.936	709 *	1·78 *	243.200

Pairs of RNAs that differ only by whether their 5′ ends originated at a promoter or at an RNAase III cleavage site in the same non-coding interval are bracketed. Asterisks identify late RNAs that appear to be identifiable in gel patterns of *in vivo* and *in vitro* transcripts of intact T7 DNA.

[a] The previous designations are used for the early mRNAs. that is. the *0·3*. *0·7*. *1*. *1·1* and *1·3* RNAs (Studier. 1973*b*). Other RNAs produced in the presence of RNAase III are referred to by the promoter from which they originated or by the RNAase III cleavage site at the left end of the RNA: the

the right end of T7 DNA (see Fig. 4). Perhaps these structures stabilize the T7 RNAs against exonucleolytic degradation from the 3′ end, and this is the primary reason why T7 mRNAs are so stable. Accumulation of stable RNAs may be an important part of the strategy by which T7 directs ribosomes to its own mRNAs, and this may be the reason why RNAase III cleavage sites are such a prominent feature of T7 transcription.

Some T7 RNAs also have the potential for base-paired structures at their 5′ ends, but this does not seem to be the rule. RNA initiated by T7 RNA polymerase at several different promoters has a polypyrimidine tract near the 5′ end that can pair with the polypurine tract that starts the RNA. In some cases a rather large structure could be formed at the 5′ end of the RNA (see, e.g. Rosa, 1981*a*). Perhaps these structures, where they occur, also contribute to the stability of T7 RNAs, or perhaps they function to direct ribosomes more efficiently to sites for initiation of protein synthesis.

All but four of the predicted T7 mRNAs code for more than one protein. Monocistronic mRNAs are predicted only for genes *1*, *1·3*, *10* and *19·5*, but, except for gene *1*, the message for each of these proteins is also found as part of one or more polycistronic mRNAs. In spite of the largely polycistronic nature of the T7 mRNAs, most amber mutations have no effect on the expression of downstream genes (Studier, 1972; unpublished results). Thus, there appear to be few polar effects at the level of transcription or translation. The only polar effect so far described (Saito & Richardson, 1981) is at the translational level: translation of gene *1·1* appears to be needed to activate translation of gene *1·2*, which follows it immediately in the same mRNA. It seems likely that synthesis of most T7 proteins is initiated independently.

7. Synthesis of T7 Proteins

(a) *Protein initiation sites*

The 50 close-packed genes of T7 actually specify 51 independent protein initiation sites: as noted previously (Dunn & Studier, 1981), gene *4* specifies two overlapping proteins, which begin at initiation sites 189 nucleotides apart and end at the same termination site. The nucleotide sequences in the mRNAs around the start sites for each of these 51 proteins, as well as those around the start sites for the five potential overlapping proteins discussed in the next section, are given in Tables 10 to 12.

The initiation codon for all but five of the 51 proteins specified by the close-packed T7 genes is AUG: the gene *2·8*, *6·3*, *7·7*, *17·5* and *19* proteins begin at GUG. Ahead of each initiation codon in the mRNA is a ribosome-binding sequence of from four to nine nucleotides, capable of uninterrupted pairing with nucleotides near the 3′ end of 16 S ribosomal RNA (Shine & Dalgarno, 1974; Steitz, 1980). All but three of these ribosome-binding sequences contain at least G-G-A-G or G-A-G-G in the mRNA: the exceptions are genes *0·6* (A-A-G-G-G-G), *0·7* (A-A-G-G-A) and *1·8* (U-A-A-G-G-A). The distance from the A (or its equivalent) in the G-G-A-G-G ribosome-binding sequence to the first nucleotide of the initiation codon ranges between seven and 13 nucleotides. The shortest interval between the last

TABLE 10

Initiation sites for synthesis of T7 early proteins

T7 protein	Pairing[a] length	Distance[b] A to ATG	Potential pairing to 16 S rRNA[c] AUUCCUCCACU-----5′
0·3	5	12	CTAATAACTGCACGAGGTAACACAAGATGGCTATGT
0·4	7	13	CGAGGAGTACGAGGAGGATGAAGAGTAATGTCTACT
0·5[d]	6	10	TTTACTTATGAGGGAGTAATGTATATGCTTACTATC
0·6	6	12	GGAATCATCAAAGGGGCACTACGCAAATGATGAAGC
0·7	5	10	AACGAACATAAAGGACACAATGCAATGAACATTACC
1	4	9	TTACTAACTGGAAGAGGCACTAAATGAACACGATTA
1·1	4	12	AGAATTACTAAGAGAGGACTTTAAGTATGCGTAACT
1·2	6	12	CAAGCGTAGCTGGGAGGGTCAGTAAGATGGGACGTT
1·3	5	11	ATTTAACCAATAGGAGATAAACATTATGATGAACAT
Average	5·3 ± 1·0	11·2 ± 1·3	

[a] The number gives the length of potential continuous base-pairing between the sequence given and the sequence at the 3′ end of 16 S ribosomal RNA, including G·U base-pairs.

[b] The number of nucleotides from the base that would be expected to pair with the central U of the 16 S rRNA sequence to the first nucleotide of the initiation codon is given.

[c] A sequence of 11 nucleotides, beginning at the 3′ end of 16 S ribosomal RNA (Shine & Dalgarno, 1974) is given at the top of the Table, with the central U underlined. The sequences of the individual initiation sites are given as the DNA sequence (T rather than U), although they are meant to represent the sequences in the messenger RNAs. The sequences are aligned so that bases that could pair with 16 S rRNA are directly below the bases they could be paired with, as shown at the top of the Table. The bases capable of uninterrupted pairing are underlined, as are the potential initiation codons.

[d] Another likely start site for the gene *0·5* protein is G-A-G-G-G-A-G-T-A-A-T-G, 2 codons ahead of the start shown in the Table.

The sequence hyphens have been omitted for clarity from this and other Tables.

paired nucleotide of the ribosome-binding sequence and the first nucleotide of the initiation codon is three nucleotides, the longest is ten.

The co-ordinate expression of the three classes of T7 proteins correlates well with the program of transcription during T7 infection, but, given the apparent stability of T7 mRNAs and the relative abruptness of the shutoff of class I and host protein synthesis (see Fig. 3), some type of regulation at the translational level seems likely. Strome & Young (1978,1980*a*,*b*) have presented evidence that T7 late mRNAs outcompete the gene *0·3* mRNA for translation both *in vivo* and *in vitro*, even though the *0·3* protein is the most actively synthesized early protein. This suggests that, among T7 mRNAs, there may be a hierarchy of ability to compete for translation, and that this may be an important factor in regulation of protein synthesis during infection. Furthermore, the rate of synthesis of individual proteins within a co-ordinately expressed group also appears to vary more widely than might be expected from relative levels of mRNA, suggesting that some of the mRNAs are translated much more efficiently than others. Among the class I proteins, those that appear to be made most efficiently are the gene *0·3*, *1·3* and perhaps gene *1* proteins; among class II proteins, the gene *2·5*, *5·5* and perhaps gene *3·5* and *5* proteins; and among class III proteins, the gene *8*, *9*, *10*, *17* and perhaps gene *15* and *16* proteins (see Fig. 3).

The nucleotide sequences around the protein initiation sites have been

TABLE 11

Initiation sites for synthesis of T7 class II proteins

T7 protein	Pairing length	Distance A to ATG	Potential pairing to 16 S rRNA ACUCCUCCACU-----5′
1·4	7	9	CTGCGTTTATAAGGAGACACTTTATGTTTAAGAAGG
1·5	8	11	CGACTCACTAAAGGAGGTACACACCATGATGTACTT
1·6	6	9	CGACTCACTAAAGGAGACACTATATGTTTCGACTTC
1·7	9	10	CAAACGAATCAAGGAGGTGTTCTGATGGGACTGTTA
1·8	6	9	TGATAAACATAAGGATAAATGTTATGCATAACTTCA
2	5	12	CTTTGGAAATCGAGAGGTCAATGACTATGTCAAACG
2·5	6	11	ACGAAACCTAAAGGAGATTAACATTATGGCTAAGAA
2·8	4	10	GCAGACGAAGACGGAGACTTCTAAGTGGAACTGCGG
3	7	7	ATATACGCAAAGGGAGGCGACATGGCAGGTTACGGC
3·5	7	12	AAGGAAAGGAAAGGAGGAAAGAAATAATGGCTCGTG
3·8	6	10	TTTGTTCGCATTGGAGGTCAAATAATGCGCAAGTCT
4A	6	10	TTTQTQGQQCTAGGAGGGAATTGCATGGACAATTCG
(*4·1*)	5	7	GTGACAACTGTGGGAGTAGTGATGGGAACTCGCTGT
4B	7	10	CGGAAACCCTCAGGAGGTAAACCAATGACTTACAAC
(*4·2*)	4	8	AACCCAGACAAAGGTAAAGCACATGAGGAAGGTCGC
4·3	6	9	CGACTCACTAAAGGAGACACACCATGTTCAAACTGA
4·5	5	10	AGTAATCAAACAGGAGAAACCATTATGTCTAACGTA
4·7	5	10	CGACTCACTATAGGAGATATTACCATGCGTGACCCT
5	5	10	CGATAATCAATAGGAGAAATCAATATGATCGTTTCT
5·3	6	9	GCCACTGATACAGGAGGCTACTCATGAACGAAAGAC
5·5	5	10	CATAAAACTATAGGAGAAATTATTATGGCTATGACA
5·7	8	10	TGCAACAGTACGGGAGGTGTTCTGATGTCTGACTAC
6	6	12	CTGAAACGAATGGGAGGATGTGTCTAATGTCTCGTG
6·3	6	10	CTTTATTGACAAGGAGATTTACCTGTGGAGACCGTA
Average[a]	6·2 ± 1·2	10·0 ± 1·2	

See the footnotes to Table 10.

[a] Averages do not include potential overlapping genes (*4·1* and *4·2*).

examined for regularities that might indicate a role in initiation for some site in addition to the ribosome-binding sequence and the initiation codon, or something that might correlate with the expression class of an mRNA or with the relative efficiency of translation within a class. There is a slight trend toward longer ribosome-binding sequences in going from class I (average length 5·3 ± 1·0) to class II (6·2 ± 1·2) to class III (7·0 ± 1·4), but there are individual exceptions to the trend. There is also a slight decrease in the distance between the ribosome-binding sequence and the initiation codon in going from class I to classes II and III (11·2 ± 1·3 : 10·0 ± 1·2 : 9·9 ± 1·8), again with individual exceptions. Thus, the length of the pairing sequence and the distance to the initiation codon do not in themselves provide sufficient information to assign individual mRNAs to one class or another.

One fairly consistent finding is that the nucleotides between the last paired nucleotide of the ribosome-binding sequence and the first nucleotide of the initiation codon are relatively rich in A and T but deficient in G. This is particularly true of the class II and III mRNAs, but again, there are individual exceptions. For all 51 sites, 72% of the nucleotides in these positions are A or T and only 9% are G; 27 sites have no G, 17 have one G, four have two G residues,

TABLE 12

Initiation sites for synthesis of T7 class III proteins

T7 protein	Pairing length	Distance A to ATG	Potential pairing to 16 S rRNA AUUCCUCCACU-----5'
0·5	4	13	GATTTAACTCTAAGAGGAATCTTTATTATGTTAACA
0·7	8	12	AAGAGATGATGGGGAGGATTGACACTATGTGTTTCT
1	6	12	AAGTCCGCATTTGGAGGTAAGAAGTGATGTCTGAGT
1·3	9	11	TGTATACTTTAAGGAGGTATAAGTTATGGGTAAGAA
1·7	7	8	AACATTTAATCAGGAGGTTATCGTGGAAGACTGCAT
2	4	11	AGCGTAAGACATGGAGACACATTTAATGGCTGAGAA
3	7	11	GTTCAACTTTAAGGAGACAATAATAATGGCTGAATC
3·5	6	10	TAACTTTAAGAAGGAGATATACATATGGCTAGCATG
4	7	9	TTTTTGCTGAAAGGAGGAACTATATGCGCTCATACG
4·3	8	8	ATTAATAAATAAGGAGGCTCTAATGGCACTCATTAG
4·7	6	11	AATACGACTACGGGAGGGTTTTCTTATGATGACTAT
5	7	11	ACAATCACGAAAGGAGGATAACCATATGTGTTGGGC
5·3	8	12	GAACCAAGACGGGGAGGTAATGAGCTATGAGTAAAA
5·5	8	9	AAGGCTACATAAGGAGGCCCTAAATGGATAAGTACG
5·7	9	7	GATTTACTTTAAGGAGGTCAAATGGCTAACGTAATT
6	8	9	ATGGACTCTCAAGGAGGTACAAGGTGCTATCATTAG
6·5	7	10	GAAAGCCAATAAGGAGTGATATGTATGGAAAAGGAT
(6·7)	6	7	TGAAAGTTAACGGGAGGCATTATGCTAGAATTTTTA
7	7	9	TAAATGGAAACAGGAGGTACACAATGAGTACGTTAA
(7·2)	8	7	GAAATCAAGTAAGGAGGCAATGTGTCTACTCAATCC
(7·3)	4	9	CTATAAGGAACTTGAGGATAACCGTGGGTACACAAC
7·5	6	11	CTTATGGAAGCTGGAGGTTTCCGTGATGGCTACTCC
Average[a]	7·0±1·4	9·9±1·8	ACCAACATAAAGGGAGGAGACTCATGTTCCGCTTAT

See the footnote to Table 10.

[a] Averages do not include potential overlapping genes (*1·7*, *7·2* and *7·3*).

and three have three G residues. This trend is even more pronounced in the 13 initiation sites (referred to above) that seem to initiate most efficiently: these mRNAs have 80% A+T and only 5% G at these positions: nine of these sites have no G, three have one G, and one has two G residues. The nucleotides preceding the ribosome-binding sequence also tend to favor A and T, in some cases very strongly, but on average not as strongly as the nucleotides between the ribosome-binding sequence and the initiation codon. No consistent correlation between nucleotide sequence ahead of the ribosome-binding sequence and translational properties of these mRNAs has yet been identified.

Analysis of codon usage in the first position past the initiation codon produced the most striking correlation with translation efficiency, a correlation pointed out by Gold *et al.* (1981). Twenty-four different codons are used as the second codon for the 51 T7 proteins: 11 codons are used once, eight are used twice, two are used three times, and one each is used four, six or eight times. The codon that is used eight times is GCU (alanine), which is the second codon for the gene *0·3*, *2·5*, *3·5*, *5·5*, *8*, *9*, *10* and *17* proteins. These eight proteins are, without exception, the most actively synthesized proteins in the cell during the time they are expressed

(Fig. 3). The only other alanine codon used, GCA, is used twice as the second codon, for the gene *3* and *12* proteins: but these proteins are not synthesized at particularly high levels. The codon used six times is UCU (serine), and the codon used four times is AUG (methionine): the appearance of these codons in the second position does not appear to correlate particularly well with either high or low efficiency of translation. Why the use of GCU as the second codon should correlate with high levels of protein synthesis is not known, but it will be interesting to see to what extent this correlation will obtain with other mRNAs that are translated in *E. coli*. It would also be interesting to determine whether translational efficiency could be affected by mutating the GCU second codons in these T7 mRNAs. No correlation between codon usage in the third position and translational properties of the T7 RNAs was detected.

(b) *Other potential T7 proteins*

We have searched the nucleotide sequence for additional sites that have properties in common with the 51 protein initiation sites associated with the 50 close-packed genes, in order to see how frequently such sequences occur, and to look for other potential T7 genes whose coding sequences might overlap one or more of the 50 close-packed genes in a different reading frame. All but three of the 51 protein initiation sites have the sequence G-G-A-G or G-A-G-G followed seven to 13 bases from the A by AUG or GUG, and it seems likely that any additional initiation sites would have a similar sequence. A computer search of the entire *l* strand of T7 DNA has identified 73 such sites in addition to the 48 expected sites. The additional 73 sites differ somewhat from the 48 expected sites: the average length of the ribosome-binding sequence is shorter (4·7 *versus* 6·4), although the average distance between the ribosome-binding sequence and the initiation codon is about the same in the two groups: almost half (35 of 73) of the additional sites have GUG as the potential initiation codon whereas only five of the 48 expected start sites do: and the nucleotides between the ribosome-binding sequences and the initiation codons of the 73 additional sites contain about equal numbers of all four bases, whereas these positions in the 48 expected sites are enriched in A and T and low in G.

Twenty-four of the 73 additional initiation sites would start a protein within the coding sequence and in the same reading frame as one of the 50 close-packed genes, and would produce proteins ranging in size from 13 to 844 amino acids. One other site would start a protein ahead of and in the same reading frame as the gene *4·5* protein, and would continue to the termination site for the gene *4·5* protein. It is not known whether any of these 25 potential in-frame proteins are made.

The remaining 48 of the 73 additional initiation sites would initiate proteins that would be read from the same sequence as one of the close-packed proteins but in a different reading frame. Thirty of the proteins that could be made from these initiation sites would be less than ten amino acids long, and only eight would contain 40 or more amino acids. Four of these eight potential proteins seem rather unlikely to be made because: (1) the initiation codon would be GUG: (2) there is only a four-nucleotide ribosome-binding sequence: and (3) the nucleotides

between the ribosome-binding sequence and the GUG are rich in G and C. This leaves four candidates for additional T7 proteins, the potential gene *4·1*, *18·7*, *19·2* and *19·3* proteins, which would range in size from 39 to 84 amino acids. The positions of the coding sequences for these four potential proteins are given in Tables 2 and 3, and the nucleotide sequences around their initiation sites are given in Tables 11 and 12.

Another way we have searched for potential T7 proteins in the nucleotide sequence is to look for long open reading frames that do not correspond to one of the close-packed T7 proteins, and to see whether they are headed by any sequence resembling a known protein initiation site. In the entire *l* strand there are 20 open reading frames 200 nucleotides or longer that do not correspond to one of the close-packed T7 genes. Four of them contain the coding sequences for the four potential overlapping proteins just discussed, and 15 have no likely initiation sites within them: the remaining open reading frame contains a plausible initiation site that would produce a protein of 112 amino acids, the potential gene *4·2* protein (Tables 2 and 11).

The five potential overlapping proteins identified in these two searches are of reasonable size relative to the known T7 proteins. Furthermore, the initiation sites for protein synthesis are generally similar to the initiation sites for the 51 close-packed T7 proteins, including the composition of the bases that lie between the ribosome-binding sequence and the initiation codon. The coding sequence for the potential gene *4·1* protein lies within the coding sequence for the *4A* protein, and the termination codon overlaps the initiation codon of the *4B* protein. Such an overlap of termination and initiation codons occurs eight times among the close-packed T7 genes. The potential gene *4·2* protein would initiate within the coding sequence of gene *4*, go past the end of the gene *4* protein, and end at the beginning of the ϕ*4·3* promoter sequence. Termination of a coding sequence at the start of a promoter sequence occurs six times among the close-packed T7 genes. The coding sequence for the potential gene *18·7* protein would lie entirely within the coding sequence of gene *18·5*, and those for the potential gene *19·2* and *19·3* proteins entirely within the coding sequence of gene *19*. The second codon for the potential gene *19·3* protein would be GCU, the second codon used for the highly expressed T7 proteins. It seems quite possible that one or more of these five potential proteins is made during T7 infection, and we have some preliminary genetic evidence that the potential gene *19·3* may in fact be expressed (unpublished results).

(c) *Frameshifting during translation*

(i) *The gene* 10A *and* 10B *proteins*

Gene *10* specifies the major capsid protein of T7. However, gene *10* amber mutants lack not only the major capsid protein (*10A*) but also a larger protein (*10B*) that is made in much smaller amounts and is also found in phage heads (Studier, 1972: see Figs 3 and 8). Both the *10A* and *10B* proteins are made *in vitro* from purified gene *10* mRNA (Fig. 8). Examination of the nucleotide sequence in the gene *10* region led us to the conclusion that the *10B* protein may be made by a shift in reading frame during translation of the *10A* protein.

The *10A* protein is predicted to begin at the AUG at nucleotide 22,966 and end at the UAA at 24,001, and we have confirmed this by determining the amino acid sequence at both ends of the *10A* protein purified from phage particles. There is no potential protein initiation site ahead of the *10A* initiation site that could produce the *10B* protein (in the way the *4A* and *4B* protein are produced), and in fact the reading frame is open ahead of the *10A* initiation codon for only five amino acids. Readthrough of the UAA termination codon could add 25 amino acids to the 344 amino acids of the *10A* protein, but a −1 frameshift during translation could add 53 amino acids, a number that seems more consistent with the relative mobilities of the *10A* and *10B* proteins upon gel electrophoresis. A shift from reading frame 1, the reading frame of the *10A* protein, to reading frame 3 at any of the 33 amino acids preceding the *10A* termination codon would allow protein synthesis to continue to the UAA at nucleotide 24,159. Such a frameshifted protein would end just ahead of the stem-and-loop that can form at the transcription termination signal for T7 RNA polymerase (Fig. 5), consistent with the close-packed arrangement of the T7 genes.

The open reading frame that could supply the additional amino acids to make the *10B* protein does not have a likely protein initiation site within it. The best potential initiation site would include a G-A-G-G (23,997) and a GUG initiation codon (24,012); however, the separation between the ribosome-binding sequence and the GUG is at the outer limit observed for T7 proteins, and the U of the GUG is preceded by five consecutive G residues and followed by two more. Initiation at this site would produce a protein of 48 amino acids, but it seems rather unlikely that this potential protein would be initiated.

To demonstrate that the *10B* protein does require a coding sequence past the *10A* termination codon at 24,001, fragments of T7 DNA were cloned in pBR322 and tested for ability to specify these two proteins during T7 infection. The cloned fragments that were tested all began ahead of the $\phi 10$ promoter and ended at various points within the coding sequence for the *10A* or *10B* protein, or past the Tϕ transcription termination site. Infection was by a gene *10* amber mutant so that any *10A* or *10B* proteins specified by the cloned fragments could be observed. The results are shown in Figure 8. When the cloned fragment contained the complete coding sequence for the $\phi 10$ mRNA (lane 5), both the *10A* and *10B* proteins were produced in normal amounts. When the cloned fragment ended within the coding sequence for *10A* (lane 6), neither the *10A* nor *10B* protein was observed at its normal position, but an equivalent amount of a protein shorter than the *10A* protein was produced. When the cloned fragment ended past the termination codon for *10A* but before the predicted termination codon for *10B* (lanes 7 and 8), the *10B* protein was missing but the *10A* protein was made in normal amounts, exactly as predicted if the *10B* protein arises by frameshifting. The cloned fragment used in lane 8 ends beyond the first termination codon that is past the *10A* termination codon in the same reading frame, ruling out the possibility that the *10B* protein arises by readthrough to the next termination codon.

Where does the shift from the *10A* to the *10B* reading frame occur? The *10A* protein is predicted to contain no tryptophan. The part of the *10B* protein past

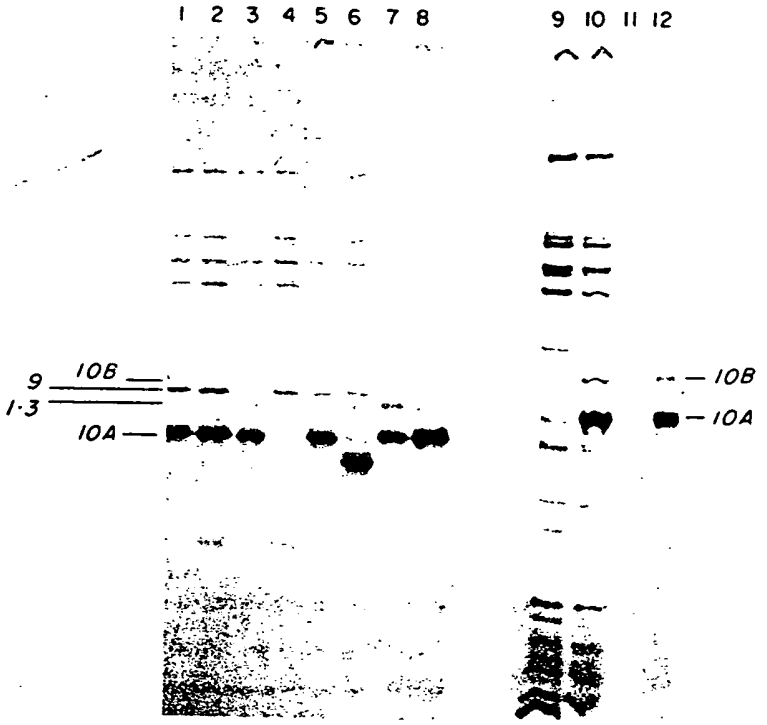


FIG. 8. Gene *10A* and *10B* proteins specified by cloned fragments of T7 DNA *in vivo* or by purified ϕ*10* RNA *in vitro*. Proteins were labeled with [^{35}S]methionine and analyzed by gel electrophoresis essentially as described in the legend to Fig. 3. Samples 1 to 8 were analyzed on a gel of 12·5% acrylamide, and samples 9 to 12 on a 10% to 20% gradient gel, both with a 5% stacking gel. Samples 1 to 10 were labeled 16 to 20 min after infection at 30°C, and samples 11 and 12 were labeled in a cell-free protein synthesis reaction for 40 min at 37°C. Infections were by: lanes 1, wild-type T7; 2, LG3, a deletion mutant lacking genes *1·1* to *1·3*; 3, *am*17 in gene *9*; 4 to 8, *am*13 in gene *10*; 9, a double mutant containing *am*17 in gene *9* and *am*13 in gene *10*; 10, *am*17 in gene *9*. The T7 mutants were described by Studier (1969) and Studier *et al.* (1979). For samples 1 to 8, the host was *E. coli* HMS174 carrying the following plasmids: lanes 1 to 4, pBR322; 5, pAR436; 6, pAR1003; 7, pAR1338; 8, pAR1061. For samples 9 and 10 the host was a *met*$^+$ derivative of *E. coli* B834 (Studier, 1981). The plasmids contain fragments of T7 DNA inserted, by means of suitable linkers, into the *Bam*HI site of pBR322 in the silent orientation, and were constructed essentially as described by Studier & Rosenberg (1981). The cloned fragments all begin at the *Cla*I site at nucleotide 22,856, ahead of the ϕ*10* promoter, and end at the following sites: pAR436 at the *Pru*II site at nucleotide 24,272, thereby including all of *10A* and *10B*; pAR1003 at the *Hae*III site at nucleotide 23,834, within *10A*; pAR1338 at the *Hae*III site at nucleotide 24,015, within *10B*; pAR1061 at the *Hae*III site at nucleotide 24,090, within *10B*. Since a gene *10* amber mutant was used to infect these strains, any *10A* and *10B* proteins observed must have been specified by the cloned fragments of T7 DNA. The positions in the gel patterns of the *10A* and *10B* proteins, and of the nearby gene *1·3* and *9* proteins, are indicated. Lanes 11 and 12 contain the [^{35}S]methionine-labeled products of protein synthesis in a cell-free system, prepared essentially as described by Goldman (1982), from *E. coli* BL15 (RNAase I$^-$ *rel*) (Studier, 1975c). The protein synthesis reaction mixture for lane 11 contained no added RNA and that for lane 12 contained purified ϕ*10* RNA that had been synthesized by purified T7 RNA polymerase (a gift from C. Fuller & C. Richardson), using pAR436 DNA as template. Rifampicin (20 μg/ml) was present to inhibit endogenous transcription. Proteins from the *in vitro* reaction mixture were precipitated by 5% (w/v) trichloroacetic acid and the pellets were washed with 90% (v/v) acetone before being dissolved in sample buffer for electrophoresis.

the *10A* termination codon would also contain no tryptophan. If the shift in reading frame occurred before nucleotide 23,976, the *10B* protein would contain one tryptophan, and if before 23,955, two tryptophans. To test whether tryptophan is found in the *10B* protein, T7 proteins were labeled with [^{3}H]tryptophan during T7 infection and were analyzed by gel electrophoresis (not shown). No label was found in the gene *5·5*, *9* or *10A* proteins, as predicted by the nucleotide sequence (Tables 14 and 15). Even after extended exposure of the autofluorogram no label was found in the position expected for the *10B* protein, indicating that the frameshift must occur between nucleotides 23,977 and 24,000. This result also provides additional evidence that the *10B* protein does not arise by readthrough of the *10A* termination codon, since such a readthrough protein would contain two tryptophans.

The *10B* protein is made at only a small percentage of the rate of the *10A* protein during T7 infection. Perhaps the frameshifting mechanism is utilized to maintain a relatively constant ratio of the two proteins that greatly favors the *10A* protein. A comparably unequal ratio of the two proteins is incorporated into phage heads. The clones that make normal *10A* but not *10B* protein may be useful in testing whether the *10B* protein may have a special role in the structure or assembly of phage heads. It might even be feasible, by deleting one nucleotide in the region where the frameshift must occur, to construct a clone that would produce only *10B* protein.

(ii) *The gene 5·5 and 5·5–5·7 fusion proteins*

Two other T7 proteins are both eliminated by an amber mutation in gene *5·5* (Studier, 1981). As with the *10A* and *10B* proteins, the shorter protein is made in much greater amounts than the longer protein, and again, the nucleotide sequence suggests that the longer protein is made by frameshifting during translation. A shift from reading frame 3 to 2 at any of the 15 amino acids preceding the *5·5* termination codon would place translation in the reading frame of the *5·7* protein and produce a *5·5–5·7* fusion protein. The *5·5* protein is predicted to contain 98 amino acids, the *5·7* protein 68 amino acids, and the *5·5–5·7* fusion protein 168 amino acids. The *5·7* protein has not been identified by gel electrophoresis, but the predicted sizes of the *5·5* and *5·5–5·7* fusion proteins agree well with the relative mobilities observed for the two proteins eliminated by the amber mutation in gene *5·5*. It should be possible to confirm that the predicted *5·5–5·7* fusion protein contains the *5·7* amino acid sequence by showing that the fusion protein is eliminated by an amber mutation in gene *5·7*. What role, if any, the *5·5–5·7* fusion protein may have during T7 infection is not known.

(iii) *The C5 deletion*

The C5 deletion of T7 shortens the gene *0·3* protein, reduces the amount of gene *1* protein made, and causes a very small amount of a gene *0·3–1* fusion protein to be made (Studier, 1973*b*). We have determined the exact location of the C5 deletion in the nucleotide sequence, which allows these effects to be explained. The C5 deletion arose by a crossover between two A-A-T-G-A-A sequences, the first located within the coding sequence for the gene *0·3* protein, at nucleotide

1231, and the second at the beginning of the coding sequence for the gene *1* protein, at nucleotide 3168. The reading frame for the *0·3* protein at the crossover sequence is AAT GAA, which leads to a termination codon three codons past the crossover sequence, thereby producing a protein of 105 instead of 116 amino acids. The ATG of the crossover sequence at 3168 is the initiation codon for the gene *1* protein. Therefore, the C5 deletion has removed the ribosome-binding sequence normally ahead of this ATG and replaced it with the sequence C-G-C-A-G-A-A-G-A-C-T-T-G-C-T-C-A-A-T-G, a much weaker ribosome-binding sequence. This is presumably why the rate of synthesis of gene *1* protein is greatly reduced.

The C5 deletion leaves the gene *0·3* and *1* reading frames out of phase, so the observed *0·3–1* fusion protein must be produced by a shift of reading frame during translation. This frameshift could occur anywhere between the TAA at 1193 and the crossover point, or between the crossover point and the TAA at 3182, a total of 17 amino acids. The reading frame must increase by one, the opposite of the shift during translation of the gene *5·5* and *10* proteins, and must occur at very low frequency, since very small amounts of a large fusion protein are observed. If the same frameshift were to occur during normal synthesis of the *0·3* protein, translation would proceed to the TAG at 1331 and produce a protein of 134 amino acids. Such a protein would probably be produced at such low levels that it would not be detected in the usual protein patterns during infection.

(iv) *The potential gene* 0·6A *and* 0·6B *proteins*

The three above examples of frameshifting led us to re-examine our previous assignment (Dunn & Studier, 1981) of coding sequences for the gene *0·6* and *0·65* proteins. The best evidence that either of these proteins is made is the observation that, during infection by certain deletion mutants of T7 (D11, C24 and C93), a new protein appears that has a size consistent with its being a fusion protein that initiated at the *0·6* initiation site and ended at the *0·7* termination codon (Simon & Studier, 1973; Studier *et al.*, 1979). (One direct observation of a protein that potentially could have been the *0·6* protein (Studier *et al.*, 1979) is difficult to interpret in the light of subsequent nucleotide sequence information, and may have been artifactual.) Two of the deletions, C24 and C93, have one end within the sequence of the potential *0·65* protein, but the size of the observed fusion protein seems substantially larger than that predicted for a *0·65–0·7* fusion protein. The ribosome-binding sequence of the potential initiation site for the *0·65* protein is predicted to be A-G-G-U, different from the ribosome-binding sequence for any other T7 protein.

It now seems possible to us that gene *0·6* may represent another example of a gene that produces two proteins by frameshifting during translation. A protein predicted to contain 53 amino acids, *0·6A*, would be produced by initiation at the ATG at nucleotide 1636 and termination at the TGA at 1795. A shift from reading frame 1 to 2 between nucleotides 1769 and 1795, an interval of eight or nine amino acids, would allow translation to continue to the TAG at 1970 and produce a protein of 111 amino acids, *0·6B*. Frameshifting would provide an explanation for the sizes of the fusion proteins generated by the C24 and C93 deletions. However, readthrough of the TGA at 1795 in the same reading frame

could produce a protein of 120 amino acids and also explain the sizes of the fusion proteins. At present, it is not possible to be certain just what proteins are specified by this region of the DNA. In the Figures and Tables of this paper, it is assumed that the frameshift mechanism is used, and that the *0·6A* and *0·6B* proteins are made.

(v) *Frameshifting in general*

Of the four established or potential examples of frameshifting during translation of T7 mRNAs, one was induced or revealed by a deletion, but the other three could well be part of normal processes or functions during T7 infection. In each case, pairs of overlapping proteins would be produced in a fixed ratio that greatly favors the smaller protein. Frameshifting has recently been found to be involved in a somewhat different way in the expression of the lysis gene of small RNA phages (Kastelein *et al.*, 1982), and appears to occur naturally at a frequency that may depend on the nucleotide sequence in the region of the frameshift and also on the physiological state of the cells (Beremand & Blumenthal, 1979; Atkins *et al.*, 1979; Fox & Weiss-Brummer, 1980; Roth, 1981). It seems quite possible that frameshifting may be used rather generally for control purposes.

The frameshifts in genes *5·5* and *10* are in the −1 direction, whereas the frameshift in the C5 deletion mutant and that proposed for gene *0·6* are in the +1 direction. No particularly interesting homologies are apparent between the nucleotide sequences in the two regions where the +1 frameshifts might occur. However, a striking homology is found between the nucleotide sequences in the two regions where the −1 frameshifts could occur: the sequence U UUC AAA occurs in the same reading frame in the frameshifting region for both the gene *5·5* and *10* proteins. This homology suggests that the U UUC AAA sequence might be involved in the frameshift, and the sequence itself suggests a possible mechanism: perhaps phenylalanine transfer RNA, which reads both UUC and UUU, occasionally slips back one base at this sequence. Slippage of phenylalanine tRNA has been proposed as one possible explanation for two examples of *in vivo* frameshifting in a yeast mitochondrial protein (Fox & Weiss-Brummer, 1980). The *φ10* mRNA is easy to isolate (see Fig. 7), and both the *10A* and *10B* proteins are made from it by *in vitro* protein-synthesizing systems from *E. coli* (Fig. 8), so it may be possible to study the frameshifting reaction *in vitro*.

(d) *Termination codons*

If frameshifting occurs during translation of the gene *0·6* and *10* proteins, the proteins made by the 50 close-packed T7 genes would terminate at 52 different sites. The preferred termination codon is UAA, which is used 30 times, followed by UGA, 16 times, and UAG, six times. There are eight instances in which a termination codon overlaps the initiation codon of the next protein, four times in the sequence U-A-A-U-G, three times as U-G-A-U-G, and once as A-U-G-A. There are six cases in which a termination codon constitutes the first three nucleotides of the conserved sequence of a promoter for T7 RNA polymerase. The next nucleotide past the termination codon is usually U (33 times) but rarely A (once).

TABLE 13

Predicted amino acid compositions of T7 early proteins

Protein	Total	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
0·3	116	10	4	6	17	0	2	17	2	4	9	10	2	6	3	2	4	4	1	8	5
0·4	50	6	1	4	1	1	3	4	2	0	1	6	2	2	2	0	4	4	1	3	3
0·5	47	7	1	0	0	1	0	0	8	0	6	7	3	2	2	0	2	3	0	2	3
0·6A	53	2	6	3	2	1	1	2	4	3	5	2	5	4	1	2	2	4	0	2	2
0·7	359	36	28	20	28	8	9	29	22	11	21	28	24	14	15	8	13	13	7	8	17
1	883	100	40	37	41	13	32	65	55	21	52	71	63	24	34	37	48	49	18	23	60
1·1	42	2	8	4	2	0	1	3	2	1	0	1	7	2	2	0	3	3	1	0	0
1·2	84	6	5	5	12	2	2	2	3	3	4	10	5	2	4	0	3	5	3	6	2
1·3	359	18	13	18	22	6	11	32	28	9	20	33	28	14	16	14	16	17	9	13	22

TABLE 14

Predicted amino acid compositions of T7 class II proteins

Protein	Total	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
1·4	51	11	0	1	0	3	1	0	2	0	3	7	3	1	2	1	1	1	2	3	9
1·5	29	2	0	0	4	2	0	0	2	2	2	5	0	4	0	2	1	0	0	1	2
1·6	86	6	11	3	3	1	0	3	2	2	8	7	8	1	6	5	5	4	0	1	10
1·7	105	14	8	11	16	9	5	10	12	5	12	16	18	5	5	12	5	9	1	7	6
1·8	48	2	2	1	5	2	1	4	1	1	1	3	4	2	3	2	6	3	3	2	0
2·5	231	21	6	8	20	3	3	26	20	2	7	10	25	4	9	13	12	10	3	11	18
2·8	139	4	11	9	5	7	4	9	12	5	6	8	10	5	0	4	15	8	3	4	10
3	148	6	8	2	7	1	2	14	13	4	10	11	22	0	8	9	10	5	3	6	7
3·5	150	13	10	5	9	3	7	9	14	7	6	10	11	2	6	4	9	6	4	3	12
3·8	121	8	13	8	5	0	4	7	8	4	5	10	10	3	4	6	7	5	5	5	4
4A	566	37	27	25	41	13	17	43	54	15	27	44	38	17	19	12	46	28	10	15	38
4B	503	35	25	21	34	9	17	40	48	12	26	42	33	16	16	9	37	25	9	13	36
(4·1)	39	2	4	3	1	1	3	1	4	0	0	5	2	1	0	1	3	5	0	1	2
(4·2)	112	7	5	3	6	1	5	1	3	3	5	18	6	1	6	6	10	11	2	4	9
4·3	70	11	6	4	2	0	5	6	1	0	2	8	7	2	3	0	5	3	0	1	4
4·5	88	11	8	8	6	0	3	4	2	3	5	6	6	3	0	0	5	8	2	1	7
4·7	135	9	6	2	7	0	3	8	11	5	8	6	15	4	2	3	7	12	4	6	17
5	704	60	38	22	40	10	25	63	60	24	36	61	52	14	25	30	25	32	18	25	44
5·3	118	7	4	3	9	3	4	6	11	1	3	8	14	4	4	1	5	11	1	5	14
5·5	98	7	3	1	6	1	5	7	5	3	4	5	11	7	7	2	7	6	0	1	10
5·7	168	7	5	4	1	3	2	2	7	1	3	5	4	1	2	1	8	5	1	2	4
6	347	25	16	15	31	8	10	33	21	4	25	25	27	7	16	12	16	17	13	10	16
6·3	37	2	1	0	2	2	0	1	4	1	4	7	2	1	0	0	2	2	2	1	3

TABLE 15

Predicted amino acid compositions of T7 class III proteins

Protein	Total	Ala A	Arg R	Asn N	Asp D	Cys C	Gln Q	Glu E	Gly G	His H	Ile I	Leu L	Lys K	Met M	Phe F	Pro P	Ser S	Thr T	Trp W	Tyr Y	Val V
6·5	84	7	5	6	6	1	5	7	6	2	7	10	2	2	2	4	3	2	0	4	3
6·7	87	6	5	2	5	1	2	7	9	0	3	2	11	2	3	4	13	6	0	0	6
7	132	7	8	1	9	6	1	12	9	7	5	9	9	3	8	2	9	5	3	6	13
7·3	98	13	4	1	2	0	5	8	14	0	3	3	16	1	0	3	7	4	0	0	14
7·7	130	2	9	5	5	3	3	9	12	6	9	10	10	3	0	5	10	12	3	5	9
8	535	57	29	17	31	3	34	45	37	1	30	56	24	20	10	24	33	27	3	22	32
9	306	35	20	15	22	0	12	45	23	4	16	16	9	6	9	9	21	15	0	11	18
10A	344	51	16	15	18	2	11	20	33	7	17	28	21	9	10	9	15	23	0	10	20
11	196	13	16	13	16	2	6	16	14	0	13	15	2	7	7	5	16	10	2	12	11
12	793	45	41	62	47	5	36	35	61	12	51	56	32	10	41	36	65	50	17	38	53
13	138	8	7	3	8	3	4	8	11	4	8	10	10	4	10	5	7	12	4	5	7
14	195	34	11	6	6	1	20	11	13	0	17	9	12	9	3	3	20	10	1	4	5
15	746	80	53	33	62	0	49	56	43	3	36	53	57	31	20	25	42	39	9	22	33
16	1318	143	74	61	81	0	32	93	116	21	61	112	95	49	50	53	82	78	9	33	75
17	552	52	34	49	36	3	29	23	50	12	26	33	24	7	23	5	31	45	15	18	37
17·5	67	9	1	4	4	0	1	3	4	0	5	6	6	2	4	1	3	2	2	2	8
18	89	6	4	3	10	0	5	6	6	3	5	11	7	4	2	2	3	4	0	3	2
18·5	143	14	14	2	14	1	7	9	10	2	7	15	11	3	3	2	8	7	4	2	8
(18·7)	82	4	6	2	1	2	2	6	3	0	4	14	5	4	2	5	10	5	1	1	5
19	585	44	35	22	44	5	22	30	38	11	35	62	32	17	21	28	33	28	8	24	37
(19·2)	84	4	7	1	0	5	5	0	4	3	1	17	1	3	0	2	9	13	1	2	6
(19·3)	56	1	9	2	0	1	1	1	1	0	1	5	1	0	2	3	9	7	1	2	9
19·5	49	3	4	1	2	3	0	1	2	1	1	12	0	1	2	0	6	3	0	2	5

TABLE 16

Predicted amino acid sequences of T7 early proteins

```
0.3     AMSAMTYNNV FDMAYEMLKE NIRYDDIRDT DDLMDAIHMA RDNAVPHYYA DIFSVMASEG IQLEFEDSGL MPDTKDVIRI LQARIYEQLT IDLMEDAEDL
        LNEYLEEVEE YEEDEE

0.4     STTNVDYGLT AQTVLFYSDM VRCGFNWSLA MRQLKELYEN NKAIALESRE

0.5     MYMLTIGLLT RLGLAVGASF GKALGVAVGS YFTACIIIGI IKGALRK

0.6A    MMKHYVMPIH TSNGATVCIP DGFAMKQRIE RLKRELRINR KINKIGSGYD RTH

0.7     MNITDINNAI DAIKALPICE LDKRQGMLID LLVEMVNSET CDGELTELND RLEMQDMWTT LKCLTDQRGF KMLGNCMFSA AYSMPLLPNR VIKVGFKKED
        SGRAYTAFCR MYDGRPGIPN VYDVQRMAGC YTVVLDRLKD CERFNNDAHY KYRETRSDII QCNSDEMDEL TGMDGEFVET CKLIRVFFEG IASFDMHSGN
    200 IMFSNGDVPY ITDPVSFSQK KDGGAFSIDP EELIKEVEEV ARQKEIDRAK ARKERMEGRL EARRFKRRNR KARKAMKAKR ERMLRAWRNA ERQERRNHEV
        AVDVLGRTNN RMLKVNMFSG DFKRLEERIA LMWRNADRMA IANQLTLNID KQLDAMLMG

1       MNTINIAKND FSDIELAAIP FNTLADHYGE RLAREQLALE HESYEMGEAR FRKMFERQLK AGEVRDNAAA KPLITTLLPK MIARINDWFE EVKAKRGKRP
        TAFQFLQEIK PEAVAYITIK TTLACLTSAD NTTVQAVASA IGRAIEDEAR FGRIRDLEAK HFKKNVEEQL NKRVGHVYKK AFMQVVEADM LSKGLLGGEA
    200 WSSWHKEDSI MVGVRCIEML IESTGMVSLH RQNAGVVGQD SETIELAPEY AEAIATRAGA LAGISPMFQP CVVPPKPWTG ITGGGYWANG RRPLALVRTH
        SKKALMRYED VYMPEVYKAI NIAQNTAWKI NKKVLAVANV ITKWKHCPVE DIPAIEREEL PMKPEDIDMN PEALTAWKRA AAAVYRKTRL RSLAVSALSS
    400 CLSKPISLLT IRPSGSLTTW TGAVRVYAVS MFNPQGNDMT KGRLTLAKGK PIGKEGYYML KIHGANCAGV DKVSFPERIK FIEENHENIM ACAKSPLENT
        WWREQDSPFC FLAFCFEYAG VQHHGLSYNC SLPLAFDGSC SGIQHFSAML RDEVGGRAVN LLPSETVQDI YGIVAKKVNE ILQRDAINGT DNEVVTVTDE
    600 NTGEISEKVK LGTKALAGDW LAYGVTRSVT KRSVMTLAYG SKEFGFRQDV LEDTIQPAID SGKGLMFTQP NQAAGYMAKL IWESVSVTVV AAVEAMNWLK
        SAAKLLAAEV KDKKTGEILR KRCAVHWVTP DGFPVWQEYK KPIQTRLNLM FLGQFRLQPT INTNKDSEID AHKQESGIAP NFVHSQDGSH LRKTVVWAHE
    800 KYGIESFALI HDSFGTIPAD AANLFKAVRE TMVDTYESCD VLADFYDQFA DQLHESQLDK MPALPAKGNL NLRDILESDF AFA

1.1     MRNFEKMTKR SWRNARDFEA TKGRKLNKTK RDRSHKRSWE GQ

1.2     GRLYSGNLAA FKAATNKLFQ LDLAVIYDDW YDAYTRKDCI RLRIEDRSGN LIDTSTFYMH DEDVLFNMCT DWLNMMYDQL KDMK

1.3     MMNIKTNPFK AVSFVESAIK KALDNAGYLI AEIKYDGVRG NICVDNTANS YWLSRVSKTI PALEMLNQFD VRWKRLLNDD RCFYKDGFML DGELMYKGVD
        FNTGSGLLRT KWTDTKNQEF HEELPVEPIR KKDKVPFKLM TGNLHIKLYA ILPLNIVESG EDCDVMTLLM QEMVKNMLPL LQEYFPEIEM QARESYEVYD
    200 MVELQQLYEQ KRREGHEGLI VKDPMCIYKR GKKSGHWKMK PENEADGIIQ GLVMGTKGLA NEGKVIGFEV LLESGRLVNA TNISRALMDE FTETVKEATL
        SQHGFFSPYG IGDNDACTIN PYDGWRCQIS YMEETPDGSL RHPSFVMFRG TEDNPQEKM
```

(e) *Amino acid sequences and compositions of T7 proteins*

The predicted amino acid sequences and compositions of the T7 proteins are given in Tables 13 to 18, and the calculated molecular weights are given in Table 4. Enzymatic and structural functions are known for many of these proteins (Table 4). The amino acid sequences should be useful for determining the structure and biochemical interactions of the T7 proteins, and comparisons of the amino acid sequences of T7 proteins and unrelated proteins that have similar functions might also be informative.

The nucleotide sequence predicts that some large T7 proteins lack certain amino acids. Reeve (personal communication) has verified that the gene *9*, *15* and *16* proteins lack cysteine, and we have verified that the gene *5·5*, *9* and *10* proteins lack tryptophan.

Knowing the locations of the coding sequences for the T7 proteins makes it possible to analyze the frequency of codon usage. Analysis of specifically labeled proteins by gel electrophoresis, as in Figure 3, can provide estimates of the relative rates of synthesis of the different T7 proteins, so correlations between level of expression and codon usage can be looked for. Superficial examination of the frequencies of codon usage has not revealed any dramatic correlations, but perhaps a more sophisticated analysis will be more revealing.

TABLE 17

Predicted amino acid sequences of T7 class II proteins

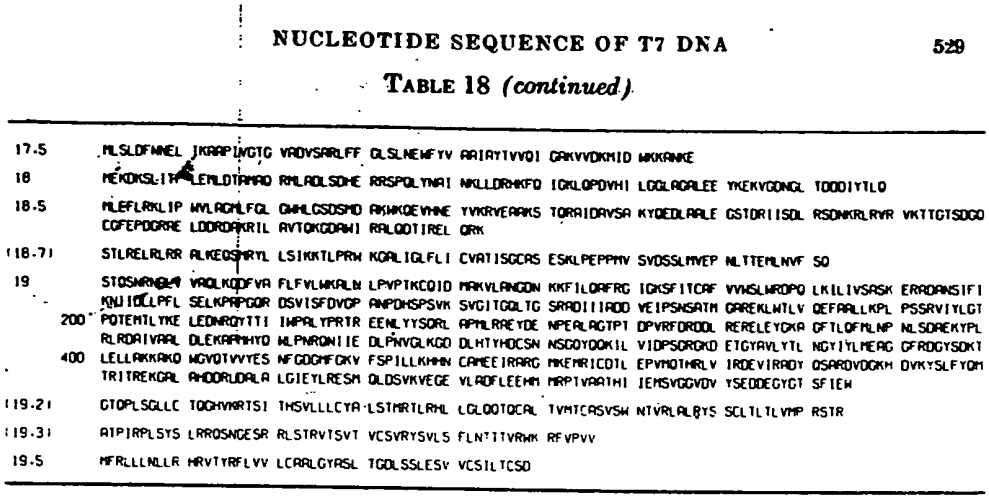
1.4		MFKKVDKFLA ALAAILTLAY ILAVYPQVAL VVVGACYLAA VCACVWSIVN W
1.5		MMYLMPLLIV IVGCLALHCS DDDMPDGHA
1.6		MFRLHYNKSV KNFTVRRADR SIVCASERRA KIPLIGNTVP LAPSVHIIIT RGDFEKAIDK KRPVLSVAVT RFPFVRLLLK RIKEVF
1.7		GLLDGEAWEK ENPPVDATGC IACLEKDDRY PHTCNKGAND MTEREQEM!! KLIDNNEGRF DDLNGCGILC SNVPCHLCPA NNDDK!ILCE IRAMDPRKPH
		LNKPEVTPTD DDPSAETIEG VTKPSHYMLF DDIEAIEVIA RSMTVEDFKG !CFGNILKYR LRAGKKSELA YLEKDLAKAD FYKELFEKHK DKCYA
1.8		MHNFKSIPPF DSLSDDFISC SEWCRKMWEE TFDDAYIKLY ELWKSRGD
2		SNVNTGSLSV DNKKFWATVE SSEHSFEVPI YAETLDEALE LAEWQYVPAG FEVTRVRPCV APK
2.5		AKKIFTSALG TAEPYAYIAK PDYGNEERGF GNPRGVYKVD LTIPNKDPRC QRMVDEIVKC HEEAYAAAVE EYEANPPAVA RGKKPLKPYE GDMPFFDNGD
		GTTTFKFKCY ASFQDKKTKE TKHINLVVVD SKGKKMEDVP IIGGGSKLKV KYSLVPYKWN TAVGASVKLQ LESVMLVELA TFGGGEDDWA DEVEENGYVA
	200	SGSAKASKPR DEESWDEDDE ESEEADEDGD F
2.8		MELREKILER IKVTSSGCWE WQGATNNKGY GQVWCSNTGK VVYCHRVMSN APKGSTVLHS CDNPLCCNPE HLSIGTPKEN STDMVNKGRS HKGYKLSDED
		VMAIMESSES NVSLARTYGV SQQTICDIRK GRRHGRLRR
3		AGYGAKGIRK VGAFRSGLED KVSKQLESKG IKFEYEEWKV PYVIPASNHT YTPDFLLPNG IFVETKGLWE SDDRKKHLLI REQHPELDIR IVFSSSRTKL
		YKGSPTSYGE FCEKHGIKFA DKLIPAEWIK EPKKEVPFDR LKRKGGKK
3.5		ARVQFKQRES TDAIFVHCSA TKPSQNVGVR EIRQWHKEQG WLDVGYHFII KRDGTVEAGR DEMAVGSHAK GYNHNSIGVC LVGGIDDKGK FDANFTPAQM
		QSLRSLLVTL LAKYEGAGLR AHHEVAPKAC PSFDLKRWWE KNELVTSDRG
3.8		MRKSYKQFYK APRRHIQVWE AANGPIPKGY YIDHIDGNPL NDALDNLRLA LPKENSWNMK TPKSNTSGLK GLSWSKEREM WRGTVTAEGK QHNFRSRDLL
		EVVAWIYRTR RELHGQFARF R
4A		MDNSHDSDSV FLYHIPCDNC GSSDGNSLFS DGHTFCYVCE KWTAGNEDTK ERASKRKPSG GKPMTYNVWN FGESNGRYSA LTARGISKET CQKAGYWIAK
		VDGVMYQVAD YRDQNGNIVS QKVRDKDKNF KTTGSHKSDA LFGKHLWNGG KKIVVTEGEI DMLTVMELQD CKYPVVSLGH GASAAKKTCA ANYEYFDQFE
	200	QIILMFDMDE AGRKAVEEAA QVLPAGKVRV AVLPCKDANE CHLNGHDREI MEQVWNAGPW IPDGVVSALS LRERIREHLS SEESVGLLFS GCTGINDKTL
		GARGGEVIMV TSGSGMGKST FVRQQALQWG TAMGKKVGLA MLEESVEETA EDLIGLHNRV RLRQSDSLKR EIIENGKFDQ WFDELFGNDT FHLYDSFAEA
	400	ETDRLLAKLA YMRSGLGCDV IILDHISIVV SASGESDERK MIDNLMTKLK GFAKSTGVVL VVICHLKNPD KGKAHEEGRP VSITDLRGSG ALRQLSDTII
		ALERNQQGDM PNLVLVRILK CRFTGDTGIA GYMEYNKETG WLEPSSYSGE EESHSESTDW SNDTDF
(4.1)		GTRCSLTDTR SATYARSGLL VMKTLKRGLQ NGNPQEVNQ
(4.2)		MRKVAPFLLL TYVVLAHYAN YLILLLPLSV ISKAICLTLS SFVFSSAALL VILVSLATWN TTRKPDGLNH QVTQGKKSHT QSQQTGPTTL TSDRILDDFP
		QDYEKFRWRV PF
4.3		MFKLIKKLGQ LLVRMYNVEA KRLNDEARKE ATQSRALAIR SNELADSAST KVTEAARVAN QAQQLSKFFE
4.5		SNVAETIRLS DTADQWNRRV HINVRNGKAT MVYRWKDSKS SKNHTQRMTL TDEQALRLVN ALTKAAVTAI HEAGRVNEAM AILDKIDN
4.7		MRDPKVIQAE IAKLEAELED VKYHEAKTRS AVHILKNLGW TWTRDTGWKK PEVTKLSHKV FDKDTMTHIK AGDWVKVDMG VVGGYGYVRS VSGKYAQVSY
		ITGVTPRGAI VADKTNMIHT GFLTVVSYEE IVKSR
5		MIVSDIEANA LLESVTKFHC GVIYDYSTAE YVSYRPSDFG AYLDALEAEV ARGGLIVFHN GHKYDVPALT KLAKLQLNRE FHLPRENCID TLVLSRLIHS
		NLKDTDMGLL RSGKLPGKRF GSHALEAWGY RLGEMKGEYK DDFKRMLEEQ GEEYVDGMEW WNFNEEMMDY NVQDVVVTKA LLEKLLSDKH YFPPEIDFTD
	200	VGYTTFWSES LEAVDIEHRA AWLLAKQERN GFPFDTKAIE ELYVELAARR SELLRKLTET FGSWYQPKGG TEMFCHPRTG KPLPKYPRIK TPKVGGIFKK
		PKNKAQREGR EPCELDTREY VAGAPYTPVE HVVFNPSSRD HIQKKLQEAG WVPTKYTDKG APVVDDEVLE GVRVDDPEKQ AAIDLIKEYL MIQKRIGQSA
	400	EGDKAWLRYV AEDGKIHGSV NPNGAVTGRA THAFPNLAQI PGVRSPYGEQ CRAAFGAEHH LDGITGKPWV QAGIDASGLE LRCLAHFMAR FDNGEYAHEI
		LNGDIHTKNQ IAAELPTRDN AKTFIYGFLY GAGDEKIGQI VGAGKERGKE LKKKFLENTP AIAALRESIQ QTLVESSQWV AGEQQVKWKR RWIKGLDGRK
	600	VHVRSPHAAL NTLLQSAGAL ICKLWIIKTE EMLVEKGLKH GWDGDFAYMA WVHDEIQVGC RTEEIAQVVI ETAQEAMRWV GDHWNFRCLL DTEGKMGPNW
		AICH
5.3		MNERHLTGAA SEMLVAYKFT KAGYTVYYPM LTQSKEDLVV CKDGKFSKVQ VKTATTVQTN TGDAKQVRLG GCGRSEYKDG DFDILAVVVD EDVLIFTWDE
		VKGKTSMCVG KRNKGIKL
5.5		AMTKKFKVSF DVTAKMSSDV QAILEKDMLH LCKQVGSGAI VPNGKQKEMI VQFLTHGMEG LMTFVVRTSF REAIKDMHEE YADKDSFKQS PATVREVF
5.7		SDYLKVLQAI KSCPKTFQSN YVRNNASLVA EAASRGHISC LTTSGRNGGA WEITASGTRF LKRMGGCV
6		SRQLVTIPRD VWNDIQGYID SLEREWDSLK NQLMEAQEYV AELEEKLNGT LQLKQFYELR EGGDQKGILV MQGQMLVFQA MSAAEFDASW EEEIWHRCCD
		MAKARQILED SIKSYETRKK AWAGAPIVLA FTDSVNWRKE LVDPNYKANR KAVKKPVGYF EFLDALFERE EFYCIREPML EGDDVMGVIA SNPSAFGARK
	200	AVIISCDKQF KTIPNCDFLW CTTGNILTQT EESAQWWHLF QTIKGQITQG YSGIAQAGDT AEDFLNNPFI TEPKTSVLKS GKNKGQEVTK WVKRDPEPHE
		TLWQCIKSIG AWAGMTEEDI IKQGQMAAIL AFNEYNFIDK EIYLWRP
6.3		METVAYIGLG LCVLGVCLIS WGLWDLARII KSLHDTK

TABLE 18

Predicted amino acid sequences of T7 class III proteins

```
6.5     MLTPINQLLK NPNQIPQVPR ATAEYLQVRF NYAYLEASGH IQLMRANGCS EAHILGFIQG LQYASNVIDE IELRKEQLRD QGED

6.7     QFSPKIKTPK MDTNQIRAVE PAPLTQEVSS VEFGGSSDET QTEGTEVSGR KQLKVERDDS VAKSKASGNG SARMKSSIRK SAFGGKK

7       SEFTCVEAKS RFRAIRWTVE MLGLPKQFEG HFVGYSLTVD EVMQMSGCRE EYILDSTQKN VAYFAMCVSC DIMMKGDILD VTSVVINPEA DSKQLQRFLA
        KRFKYLAELH QCQMVSRCKH EGETMRVYFK EV

7.3     QKKVKKAVKK VTKSVKKVVK EGARPVKQVA GGLAGLAGGT GEAQMVEVPQ AAAQIVQVPE KEVSTEDEAQ TESGRKKARA GGKKSLSVAR SSGGGINI

7.7     MEDCIEWTGG VNSKGYGRKW VNGKLVTPHR HIYEETYGPV PTGIVVMHIC DNPRCYNIKH LTLGTPKDNS EDMVTKGRQA KGEELSKKLT ESDVLAIRSS
        TLSHRSLGEL YGVSQSTITR ILQRKTWRHI

8       AEKRTGLAED GAKSVYERLK NDRAPYETRA QNCAQYTIPS LFPKDSDNAS TDYQTPWQRV GARGLNNLAS KLMLALFPMQ TMARLTISEY EAKQLLSQPQ
        QLAKVDEGLS MVERIIMNYI ESNSYRVTLF EALKQLVVAG NVLLYLPEPE QSNYNPMKLY RLSSYVVQRD AFGNVLQMVT RDQIAFGALP EDIRKAVEGQ
    200 GGEKKADETI QVYTHIYLDE DSGEYLRYEE VEGMEVDGSQ GTYPKEACPY IPIRMVRLDG ESYGRSYIEE YLGDLRSLEN LQEAIVKMSM ISSKVIGLVN
        PAGITQPRRL TKAQTGDFVT GRPEDISFLQ LEKQADFTVA KAVSDAIEAR LSFAFMLNSA VQRTGERVTA EEIRYVASEL EDTLGGVYSI LSQELQLPLV
    400 RVLLKQLQAT QQIPELPKEA VEPTISTQLE AIGRGQDLDK LERCVTAWAA LAPMRDDPDI NLAMIKLRIA NAIGIDTSGI LLTEEQKQQK MAQQSMQMGH
        DNGAAALAQG MAQGTASPE AMAAAADSVG LQPGI

9       AESNADVYRS FGVNSAVMSG GSVEEMEQNM LALDVAARDG DDAIELASQE VETEROLYQN SDPFGQEDDE GRIQVRIGDG SEPTDVQTGE EGVEGTEGSE
        EFTPLQETPE ELVARSEQLG EHEEGFQEMI NIAAERGMSV ETIEAIQREY EENEELSAES YAKLAEIGYT KAFIDSYIRG QEALVEQYVN SVIEYAGGRE
    200 RFDALYNMLE TMNPEAAQSL DNALTNRDLA TVKAIINLAG ESRAKAFGRK PTRSVTNRAI PAKPQATKRE QFRQRSEMIK AMSQPRYRTQ ANYRRQVEQK
        VIDSNF

10A     RSMTGGQQMG TNQGKGVVAR GDKLALFLKV FGGEVLTAFA RTSVTTSRMH VRSISSGKSA QFPVLGRTQA ATLAPGENLD DKRKQIKHTE KVITIDGLLT
        RDVLIYDIED AMNHYDVRSE YTSQLGESLA MARDGAVLAE IAGLCNVESK YNENIEGLGT ATVIETQNK AALTDDVRLG KETIAALTKA RAALTKNTVP
    200 ARDRVFYCDP DSYSAILAAL MPNAANYAAL IDPEKDSIRN VMGFEVVEVP HLTAGGAGTA REGTTGDKHV FPRNKQEQNV KVAKDNVIGL FMHRSAVGTV
        KLRDLALERA RRANFDRDDI IAKYAMGHGG LRPEAAGAVV FKVE

11      MRSYDMNVET ARELSAVNDI LASIGEPPVS TLEGDAADA ANARRILNKI NRQIQSRGMT FNIEEGITLL PDVYSNLIVY SDDYLSLMST SGDSIYVNRG
        GYVYDRTSQS DRFDSGITVN IIRLRDYDEM PECFRYMIVT KASRQFNNRF FGAPEVEGVL QEEEDEARRL CMEYEMDYGG YNMLDGDAFT SGLLTR

12      ALISDSIKNL KGGISDDPDI LRYPDDGSRQ VNGWSSETEG LDKRPPLVFL NTLGDNGRLG QAPYIMLINR DEHEQTYAVF TGSGIRVFDL SGNEKDVRYP
        NGSNYIKTAN PRNDLRMVTV ADYTFIVNRN VVADKNTKSV NLPNTNPNDD QLINVRGDGY GRELIVHING KDVAKYYIPD GSQPEHVNNT DRDWLAEELA
    200 KQMRTNLSDK TVNVGDGFIH VTAPSGDDID SFTTKDGYRD QLINPVTHYR QSFSKLPPNA PNGYMVKIVG DRSKSADQYY VRYDAERKVM TETLGWNTED
        DVLMETMPHA LVRAADGNFD FKWLEKSPKS CGDVDTNPWP SFVGSSINDV FFFRNRLGFL SGENIILSRT AKYFNFYPRS IANLSDDDPI DVRVSTNRIA
    400 ILKYAVPFSE ELLIKSDEAD FVLTASGTLT SKSVELNLTT QFDVDDRARP FGIGRNYTFA SPRSSFTSIM RYTAVQDVSS VRNAEDTTSH VPNYIPNGVF
        SICDSGTENF CSVLSHGDPS KIFMYKFLYL NEELRQQSHS HWDFGENYQV LACQSISSDM YVILRNEFNT FLARISFTKN AIDLQGEPYR AFMDMKIRYT
    600 IPSGTYNDDT FTTSIMIPTI YGRNFGRGKI TVLEPDGKIT VFEQPTAGWN SDPWLRLSGN LEGRMVYIGF NINFVYEFSK FLIKDTADDG STSTEDIGRL
        QLRRAWVNYE NSGTFDITVE NQSSNWKYTM AGARLGSNTL RAGRLNLGTG QYRFPVVGNA KFNTVYILSD ETTPLNIIGC GWEGNYLRRS SGI

13      MNTIRPTKST QFEVFTPAHH DILEAKAAGI EPSFPDRSEC VTLSLYGFPL AIGGNCGDQC WFVTSQQVWR LSGKAKRKFR KLIMEYRDKM LEKYDTLWNY
        VWVGNTSHIR FLKTIGAVFH EEYTRDGDFQ LFTITKGG

14      CWAAAIPIAI SGAQAISGDN AQAKMIAAQT AAGRRDAMEI MRDTNIDNAD LSLQARSKLE EASAELTSDN MQKVQAIGSI RAAIGESMLE GSSMDRIKRV
        TEGQFIREAN MVTENYRRDY QAIFAQDLGG TDSAASQIDE IYKSEQKDKS KLQMVLDPLA IMGSSARSAY RSGAFDSKST TKAPIVAAKG IKTGR

15      SKIESALQAA QPGLSALRGG AGGMGYRAAT TQREQPRSSL LDTIGRFAKA GAQMYTAKEQ RARDLAQERS NEIIRKLTPE QRREALNNGT LLYQQDPYAM
        EALRVKTGRN AAYLVQDDVM QKIKEGVFRT REEMEEYRHS RLQEGAKVYA EQFGIDPEDV QYQRGFNGDI TERNISLYGA MDNFLSQQRD KGAIMNSRVE
    200 LNGVLQDPDM LRRPDSAQFF EKYIDNGLVT GAIPSDAQAT QLISQAFSDA SSRAGGRQFL MRVGDKKVTL NGATTTYREL IGEEQMNALM VTAQRSQFET
        DAKLNEQYRL KINSALNQED PRTAWEMLQG IKRELDKVQP DEQMTPQREM LISAQEQVQN QMNAWTKAQA KALDDSMKSM NKLDVIDKQF QKRINGEMVS
    400 TDFKDMPVNE NTGEFKHSDM VNYANKKLAE IDSMDIPDGA KDAMKLKYLQ ADSKDGAFRT AIGTMVTDAG QEWSARVING KLPERTPAMD ALRRIRNAQP
        QLIAALYPDD RELFLTMDMM DKDGIDPDVI LDRDRLTVKR SKEQRFEDDK AFESALNASK APEIARMPAS LRESARKIYD SVKYRSGNES MANEQMTKFL
    600 KESTYTFTGD DVDGDTVGVI PKNMMDVNSD PKSKEQCRDI LEEARKGIIA SNPWITNKQL TMYSQGDSIY LMDTIGDVRV RYDKELLSKV WSENQKKLEE
        KAREKALADV NKRAPIVAAT KAREAAAKRV REKRKQTPKF IYGRKE

16      MDKYDKNVPS DYDGLFDKAA DANGVSTDLL RKVAWTESRF VPTAKSKTGP LDMMDFTKAT AKALGLRVTD SPDDDRLMPE LAINAAAKGL AQLVGKFDGD
        ELKAALATNQ GESRLGNPDL SAYSKGDFAS ISEESGNYMR NLLDVAKSPM AGQLETFGGI TPKQKDIPAE VGLRGIGMKG KVTQELPEST SFDVKGIEDE
    200 ATAKPFAKDF WETHGETLDE INSRSTFFGF KNAAEAELSN SVAGMAFRAG RLDNGFDVFK DTITPTRMNS MIWIPEELEK IRTEVKNPAY INVVTGGSPE
        NLDDLIKLAN ENFENDSRAA SRGLGAKLSA GIIGAGVDPL SYVPMVGYTG KQFKLINKAL VVGRESAALN ARSEGLRTSV AGGDAQYRGA ALGGFVFGRG
    400 MSAISDAVAA QLKRSKPEAE FDNEFIGPMM RLEARETARN ANSADLSRMN IENMMFEGEH NGVPYEDLPI ERGAVVLMGG SVLSASNPIN PKTLKEFSEV
        DPEKAARGIK LAGFTEIGLK TLGSDDRDIR RVAIDLVRSP TDMDSGASGK FGATASDIME RLMGTDDRTY NDLYKAMSDA MKDPEFSTGG AKMSREETRY
    600 TITRRAALAI ERPELQKALT PSERIVRDII KRMFDTKREL MENPAIFGNT KRVSIFPESR MKGTYVPHYT DRNAKALMIQ RYGREGLDEG IARSMNSYV
        SRPEVKARVD EMLKELMGVK EVTPEMVEKY AMDKAYGISM SDDFINSSII EENIEQLVGI ENNSFLEARN LFDSDLSITM PDGQQFSVNQ LRDFDMFRIM
    800 PAYDRRVNGD IAIMGSTGKT TKELKDEILA LKAKAREGDK KTGEVHALMD TVKILTGRAR RNDDTVWETS LRAINDLGFF AKNAYMGRQN ITEIRGHIVT
        GNVRALGMGI PILRDTLTKS KPVSAKELKE LMASLFGKEV DDLIRPKARD IVGRLREATD TGPAVANIVG TLKYSTQELA ARSPMTKLLN GTTNTLLDAA
   1000 RDGMLGDVIS ATLTGKTTRH EKEGFLRGRS VTPEDMAGIK SLIKEMMVRG EDGKFTYKDK QAFSMDPRAN QLWRLADKVAPDERALRPHKV SLQDSHAFGA
        LGKNVMQFKS FTIKSLNSKF LRTFYDGYKN NRAIQARLSI ITSMGLAGGF YAMAAHVKAY ALPKEKRKEY LERRLDPTNI AHARLSRSSQ LGAPLAMVDL
   1200 VGGVLGFESS KMARSTILPK DIVKERDPNK PYTSREVMGA MGSNLLEDMP SAGFVANVGA TLMNAAGVYN SPNKATEQDF MTQLMNSTKE LVPNQPLTQQ
        LVLKIYEANG VNLRERRK

17      ANVIKTVLTY QLQGSNRDFN IPFEYLARKF VVVTLIGVDR KVLTINTDYR FATRTTISLT KAWGPRDGYT TIELRRVTST TDRLVDFTDG SILRAYDLNV
        AQIQTMHVRE EARDLTTDTI GVNNQGHLDA RGRRIVNLAN AVDDRDAVPF CQLKTMNDNS MDARNERLQF RNERETFRNQ AEGFKNESST NATNTKDMRD
    200 ETKGFRDEAK RFKNTACQYA TSAGNSASAA HDSEVNRENS ATRSANSAHL REQDRDRAER ERDKLENYNG LAGRIDKVDG TNVTMKGNIM ANGRLYMTTN
        GFDCGDYQDF FGGVTNRTSV MEWGDENDAL MTVDRREWTT AIGGNIQLVV NGQITTDGGA MTGDLKLDNG MVLQLESRSD KAHTILSKDG NRNNWTIGRG
    400 SDNNNDCTFH STVMGTTLTL KDDYAVVNKH FMVGDAVVAT DGNIQGTKMG GKWLDRTLRD SFVRKSKAWT QNWSGSADDG VSVTVSGDLR FRNIMIKCAN
        NSWNFFRTGP DGITFIRSDG GWLRFQIHSN GLGFKNIRDS RSVPNRIMVE NE
```

TABLE 18 *(continued)*



17.5	MLSLDFMNEL JKRAPLVGTQ VRDVSARLFF QLSLNEWFYV ARIAYTVVQI GRKVVDKMID WKKANKE
18	MEKDKSLIFA LEMLDTRRAD RALADLSDHE RRSPQLYNAI NKLLDRMKFD IGKLQPDVHI LGGLAGALEE YKEKVGDNGL TDDDIYTLD
18.5	MLEFLRKLIP WVLAGMLFQL QWMLGSDSMD RSWKQEVNNE YVKRVEAAKS TQRAIDRVSA KYQEDLAALE GSTDRIISDL RSDNKRLRVR VKTTGTSDGD CQFEPDGRRE LDDRDARRIL AVTQKCDAWI RALQDTIREL QRK
(18.7)	STLRELRLRR ALKEQSNRYL LSIKKTLPRW KGALIGLFLI CVATISGCAS ESKLPEPPNV SVDSSLNVEP NLTTEMLNVF SD
19	STQSMRNGLA VAQLKQDFVA FLFVLWKALN LPVPTKCQID MAKVLANGDN KKFILDAFRG IGKSFITCAF VVWSLWRDPQ LKILIVSASK ERRDANSIFI KNIIDLLPFL SELKPAPGDR DSVISFDVDP ANPDHSPSVK SVGITGQLTG SRADIIIADD VEIPSNSATM DAREKLNTLV QEFAALLKPL PSSRVIYLGT 200 POTEMILYKE LEDNRQYTTI IWPRLYPRTR EENLYYSQRL APMLRREYDE NPERLAGTPT DPVRFDRDDL RERELEYQKA GFTLQFMLNP NLSDREKYPL RLRDAIVRAL DLEKAPMHYD WLPNRQNIIE DLPNVGLKGD DLHTYHDCSN NSGDYDDKIL VIDPSSRDKD ETGYAVLYTL NGYIYLMEAG GFRDGYSDKT 400 LELLAKKAKD WGVDTVVYES NFGDDMFDKV FSPILLKHHN CAMEEIRARG MKEMRICDTL EPVMQTHRLV IRDEVIRADY QSARDVDDKH DVKYSLFYQM TRITREKGAL AHDDRLDALA LGIEYLRESM DLDSVKVEGE VLRDFLEEMM MRPTVAATHI IEMSVGGVDV YSEDDEGYDT SFIEW
(19.2)	GTDPLSGLLC TDDHVKRTSI THSVLLLCYA LSTMRTLRML LGLDDTDCAL TVMTCASVSW NTVRLALRYS SCLTLTLVMP RSTR
(19.3)	AIPIRPLSYS LRRDSNGESR RLSTRVTSVT VCSVRYSVLS FLNTITVRWK RFVPVV
19.5	MFRLLLNLLR HRVTYRFLVV LCRALGYRSL TGDLSSLESV VCSILTCSD

8. Origins of DNA Replication

The primary origin of replication of T7 DNA. that is, the preferred origin for the first replication of parental DNA, has been mapped by Tamanoi *et al.* (1980) and Saito *et al.* (1980) to the non-coding region between genes *1* and *1·1*. A secondary origin, utilized as the first origin when the primary origin is deleted, is probably associated with the ϕOL promoter for T7 RNA polymerase, near the left end of the DNA (Tamanoi *et al.*, 1980; Dunn & Studier, 1981). We have confirmed the location of the primary origin and identified other potential secondary origins by analyzing the ability of cloned fragments of T7 DNA (Studier & Rosenberg, 1981) to serve as origins of replication in a plasmid during T7 infection (unpublished results). Fragments representing all of the T7 DNA except the region around the ϕOL promoter have been tested: relatively strong origin activity is associated with the primary origin and with the ϕOR and the $\phi 13$ promoters; much weaker origin activity is associated with some other promoters.

All of the origins of T7 DNA replication identified *in vivo* contain a promoter for T7 RNA polymerase, and T7 RNA polymerase is needed for proper initiation *in vitro* (Scherzinger & Seiffert, 1975; Fischer & Hinkle, 1980; Romano *et al.*, 1981). The specificity of T7 RNA polymerase for its own promoters is an important part of the mechanism for switching transcription from host DNA to T7 DNA during infection: T7 RNA polymerase is induced, the host RNA polymerase is inactivated, and thereafter all transcription in the cell is directed to T7 DNA. Apparently, the same strategy is used to switch replication from host DNA to T7 DNA: the T7 replication complex is induced, the host replication apparatus is inactivated, and the T7 replication complex is designed so as to require an interaction of T7 RNA polymerase with one of its own promoters in order to initiate replication. This arrangement uses the specificity of T7 RNA polymerase to direct all replication in the cell to T7 DNA. However, not all promoters for T7 RNA polymerase are parts of replication origins. Knowledge of the nucleotide sequence, together with the cloning assay for origin activity, may make it possible to determine precisely what is required to define an origin of replication for T7 DNA.

9. The Ends of T7 DNA, and the Concatemer Junction

The longest stretches of T7 DNA that do not code for any proteins are at the ends of the molecule, positions 0 to 2·3, 98·0 to 98·6 and 99·0 to 100. The first and last 0·4% of the molecule contain a perfect direct repeat of 160 base-pairs. T7 DNA is replicated as concatemers, that is, long molecules containing tandemly repeated T7 genomes (Kelly & Thomas, 1969). At the junction between adjacent genomes in a concatemer, the non-coding regions that will ultimately be at the ends of the mature DNA flank a single copy of the terminal repetition (Langman *et al.*, 1978; our unpublished work): the terminal repetition of mature DNA is generated during maturation and packaging. In Figure 9, the features of the region between genes *19* and *0·3* are drawn to scale as they would be found in a concatemer junction.

As pointed out (Dunn & Studier, 1981), the non-coding region at the left end of mature T7 DNA contains several prominent features: from left to right, these include the terminal repetition: a regular array of 12 short repeated sequences: an A+T-rich region that contains the ϕOL replication origin: the A1, A2 and A3 promoters for *E. coli* RNA polymerase: the R*0·3* RNAase III cleavage site: and finally, the start of the coding sequence of gene *0·3* (see Fig. 9). Some rather similar features are found near the right end of mature T7 DNA: from right to left, prominent features include the terminal repetition; an array of 12 short repeated sequences similar to that found near the left end: the coding sequence of gene *19·5*: an A+T-rich region that contains the ϕOR replication origin; and finally, the end of the coding sequence of gene *19* (see Fig. 9). To facilitate discussion, the region of DNA occupied by the array of 12 repeated sequences

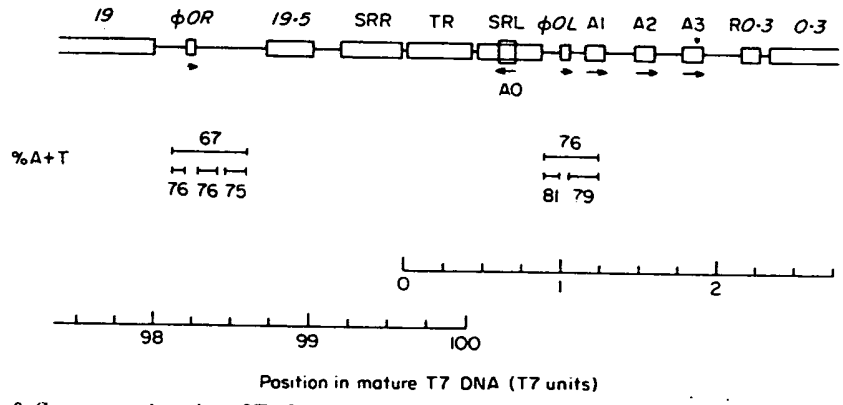


FIG. 9. Concatemer junction of T7 DNA. The arrangement of sequence elements between the coding sequences of genes *19* and *0·3* in a concatemer junction is shown. The locations of the single copy of the terminal repetition (TR) and of the arrays of short repeated sequences (SRR and SRL) near the right and left ends of mature T7 DNA are indicated. The leftward minor promoter for *E. coli* RNA polymerase (A0) lies within SRL. The positions of the strong early promoters (A1, A2 and A3) and of the first RNAase III cleavage site (R*0·3*) are also indicated. The ϕOR and ϕOL promoters for T7 RNA polymerase, and their associated A+T-rich sequences, apparently serve as origins of replication in the T7 DNA. The arrows indicate the direction of transcription from the promoters. The scale is in map units.

adjacent to the terminal repetition at the left end of mature T7 DNA will be referred to as SRL (short repeats, left end), and the similar region at the right end of mature T7 DNA will be referred to as SRR (short repeats, right end).

SRL and SRR both contain 12 copies of the sequence C-C-T-A-A-A-G, or variants of it, arranged in two sets of six. The copies are less variable in sequence and are more regularly spaced in SRL than in SRR and, in fact, the array in SRR would probably not have been identified without SRL as an example. In SRL, eight of the 12 copies are C-C-T-A-A-A-G and four copies differ from this sequence in one position; in SRR, only two of the 12 copies are C-C-T-A-A-A-G, six differ from this sequence in one position, three differ in two positions, and one differs in three positions. The spacing of the repeated sequences in SRL (beginning at nucleotide 175) is a very regular *13, 13, 12, 13, 13*, 28, *13, 13, 13, 13, 13* nucleotides: the spacing in SSR (beginning at nucleotide 39,605) is a similar but less regular *15, 13, 11, 13, 13*, 29, *12, 11, 11, 12, 12* nucleotides. Additional homologies between nucleotides adjacent to the basic repeated elements increase the length of some perfect repeats to as long as 23 base-pairs in SRL (Dunn & Studier, 1981) and as long as nine base-pairs in SRR. The longest perfect homologies between SRL and SRR are nine base-pairs. SRL occupies 164 base-pairs, from the first nucleotide of the first repeated element to the last nucleotide of the 12th, and SRR occupies 159 base-pairs. SRL begins 15 nucleotides past the end of the terminal repetition and SRR ends 14 nucleotides before the terminal repetition. The leftward A0 promoter for *E. coli* RNA polymerase lies within SRL: it occupies the interval between the two sets of six repeats, and overlaps the last two repeats in the first set; RNA chains would begin within the first set of six repeats, at nucleotide 224. No equivalent promoter has been identified within SRR.

The location of SRL and SRR relative to the terminal repetition suggests that these arrays of short repeated sequences may have a role in forming the mature ends of T7 DNA. Perhaps the nearby origins of replication associated with *φOL* and *φOR* are also involved. Maturation of the DNA is known to be associated with packaging into phage heads, and the gene *18* and *19* proteins are also required (Studier, 1972). It is interesting that all of the elements identified across the concatemer junction, including the terminal repetition itself, have an intrinsic polarity and, although elements such as SRR and SRL, or *φOR* and *φOL*, are approximately symmetrically positioned relative to the terminal repetition, the polarities have the same orientation relative to T7 DNA itself and are therefore not symmetrical about the terminal repetition. The molecular details of the maturation and packaging process are far from clear, but perhaps knowledge of the nucleotide sequence, and of the arrangement of elements around the concatemer junction, will be helpful in working them out.

10. Other Features of the Nucleotide Sequence

The nucleotide sequence given in Figure 1 for the *l* strand of T7 DNA contains 10,841 A residues, 9767 T residues, 10,291 G residues and 9037 C residues, which amounts to 27·2% A, 24·5% T, 25·8% G and 22·6% C in the *l* strand. The double-

stranded molecule contains 20.608 residues each of A and T, and 19.328 residues each of G and C, which amounts to 48·4% G+C. The molecular weight of the sodium salt of T7 DNA is calculated to be $26{\cdot}43 \times 10^6$.

Uninterrupted runs of a single base occur much less frequently in T7 DNA than would be predicted for a random sequence of nucleotides. No such runs of seven or longer occur, and only one run of six, are found in the *l* strand, which is the run of six T residues found at the transcription termination site Tϕ. In addition to this run of six, the *l* strand contains T-T-T-T-T once, C-C-C-C-C and G-G-G-G-G three times each, and A-A-A-A-A seven times.

Uninterrupted runs of pairs of nucleotides are also under-represented in T7 DNA. If we consider such runs of eight nucleotides or longer, only runs containing A or G have an approximately random distribution of lengths, and then only in the *l* strand: 126 such runs are found between eight and 14 long, plus one run of 22 (which occurs within the coding sequence of gene *3*). For other pairs of nucleotides, the number of runs eight or longer in the *l* strand is 80 of G, T: 59 of A, C: 35 of A, T: 32 of C, T: and only five of G, C: the longest uninterrupted runs are 14 of G, T: 13 of A, C: 17 of A, T: ten of C, T: and eight of G, C.

The locations of most of the runs of pairs of nucleotides is within the coding sequences for T7 proteins, as would be expected if these runs were distributed essentially at random. An exception is runs containing only A or T, where 11 of the 16 runs of ten or longer, and 19 of the 35 runs of eight or longer, are found in non-coding sequences. This distribution presumably reflects a physiological role for A+T-rich regions in opening the DNA at promoters and replication origins, and perhaps in destabilizing RNA structures ahead of initiation sites for protein synthesis. In addition, four of the runs containing only C or T are located just downstream of promoters for T7 RNA polymerase, where they can base-pair with the polypurine tract at the beginning of the RNA, and two such runs from parts of the paired structure at RNAase III cleavage sites.

All perfectly repeated sequences of ten continuous base-pairs or longer in T7 DNA have also been identified. The longest repeated sequence in the molecule is the terminal repetition, 160 base-pairs long. Because there are 17 promoters for T7 RNA polymerase, which share a high degree of homology over 23 continuous base-pairs, large numbers of relatively long repeated sequences are found. A random sequence 39,936 nucleotides long would be predicted to have an average of less than one repeated sequence 15 base-pairs long or longer (Dunn & Studier, 1981), but T7 DNA has 76 such pairs of repeated sequences. The longest perfect repeat involving promoters is 30 base-pairs, between the *ϕ10* and *ϕ13* promoters. Other than the terminal repetition and repeats between promoter sequences, the longest repeated sequences are one repeat of 23 and two of 20 within the SRL sequence near the left end of T7 DNA (see section 9), and repeats of 17 between the A2 and A3 promoters and between the R*0·7* and R*1·1* RNAase III cleavage sites, all of which have been pointed out (Dunn & Studier, 1981). Two repeats of 16 and three of 15 base-pairs are found between sequences that are not parts of promoters, and these are repeats between coding sequences in different genes. As repeats become shorter still, a larger fraction is found between sequences that are not parts of promoters. However, no other class of

large repeated sequences analogous to promoter sequences has been identified in the DNA.

We thank C. Fuller and C. Richardson for a gift of purified T7 RNA polymerase, and D. Botstein, J. Reeve and M. Rosa for communicating unpublished results. We thank K. Thompson for extensive computer analyses. N. Alonzo and M. Elzinga for help in determining amino acid sequences at the ends of the gene *10A* protein. A. Rosenberg for valuable help in mapping restriction fragments and for constructing recombinant plasmids, and W. Crockett and B. Lade for dedicated assistance in the sequencing effort. This research was carried out at Brookhaven National Laboratory under the auspices of the United States Department of Energy.

REFERENCES

Allet, B., Roberts, R. J., Gesteland, R. F. & Solem, R. (1974). *Nature (London)*. **249**, 217–221.
Atkins, J. F., Gesteland, R. F., Reid, B. R. & Anderson, C. W. (1979). *Cell*. **18**, 1119–1131.
Barrell, B. G., Air, G. M. & Hutchison, C. A. III (1976). *Nature (London)*. **264**, 34–41.
Beremand, M. N. & Blumenthal, T. (1979). *Cell*. **18**, 257–266.
Boothroyd, J. C. & Hayward, R. S. (1979). *Nucl. Acids Res.* **7**, 1931–1943.
Carter, A. D. & McAllister, W. T. (1981). *J. Mol. Biol.* **153**, 825–830.
Carter, A. D., Morris, C. E. & McAllister, W. T. (1981). *J. Virol.* **37**, 636–642.
Chamberlin, M. & Ring, J. (1973). *J. Biol. Chem.* **248**, 2235–2244.
Dayhoff, M. O. (1972). *Atlas of Protein Sequence and Structure*, vol. 5, p. D-4, National Biomedical Research Foundation, Silver Spring.
Delius, H., Westphal, H. & Axelrod, N. (1973). *J. Mol. Biol.* **74**, 677–687.
Dunn, J. J. (1976). *J. Biol. Chem.* **251**, 3807–3814.
Dunn, J. J. & Studier, F. W. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 1559–1563.
Dunn, J. J. & Studier, F. W. (1975). *Brookhaven Symp. Biol.* **26**, 267–276.
Dunn, J. J. & Studier, F. W. (1980). *Nucl. Acids Res.* **8**, 2119–2132.
Dunn, J. J. & Studier, F. W. (1981). *J. Mol. Biol.* **148**, 303–330.
Dunn, J. J., Buzash-Pollert, E. & Studier, F. W. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 2741–2745.
Fischer, H. & Hinkle, D. C. (1980). *J. Biol. Chem.* **255**, 7956–7964.
Fox, T. D. & Weiss-Brummer, B. (1980). *Nature (London)*. **288**, 60–63.
Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Singer, B. S. & Stormo, G. (1981). *Annu. Rev. Microbiol.* **35**, 365–403.
Goldman, E. (1982). *J. Mol. Biol.* **158**, 619–636.
Golomb, M. & Chamberlin, M. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 760–764.
Grachev, M. A. & Pletnev, A. G. (1981). *FEBS Letters*, **127**, 53–56.
Hausmann, R. & Gomez, B. (1967). *J. Virol.* **1**, 779–792.
Kassavetis, G. A. & Chamberlin, M. J. (1979). *J. Virol.* **29**, 196–208.
Kastelein, R. A., Remaut, E., Fiers, W. & van Duin, J. (1982). *Nature (London)*. **295**, 35–41.
Kelly, T. J. Jr & Thomas, C. A. Jr (1969). *J. Mol. Biol.* **44**, 459–475.
Kiefer, M., Neff, N. & Chamberlin, M. J. (1977). *J. Virol.* **22**, 548–552.
Kramer, R. A., Rosenberg, M. & Steitz, J. A. (1974). *J. Mol. Biol.* **89**, 767–776.
Langman, L., Paetkau, V., Scraba, D., Miller, R. C. Jr, Roeder, G. S. & Sadowski, P. D. (1978). *Canad. J. Biochem.* **56**, 508–516.
Marrs, B. L. & Yanofsky, C. (1971). *Nature New Biol.* **234**, 168–170.
Maxam, A. M. & Gilbert, W. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 560–564.
Maxam, A. M. & Gilbert, W. (1979). In *Methods in Enzymology* (Grossman, L. & Moldave, K., eds), vol. 65, pp. 499–559, Academic Press, New York.
McAllister, W. T. & Carter, A. D. (1980). *Nucl. Acids Res.* **8**, 4821–4837.

McAllister, W. T. & McCarron, R. J. (1977). *Virology*. **82**, 288–298.
McAllister, W. T. & Wu, H.-L. (1978). *Proc. Nat. Acad. Sci., U.S.A.* **75**, 804–808.
McAllister, W. T., Morris, C., Rosenberg, A. H. & Studier, F. W. (1981). *J. Mol. Biol.* **153**, 527–544.
McConnell, D. J. (1979). *Nucl. Acids Res.* **6**, 525–544.
McDonell, M. W., Simon, M. N. & Studier, F. W. (1977). *J. Mol. Biol.* **110**, 119–146.
Millette, R. L., Trotter, C. D., Herrlich, P. & Schweiger, M. (1970). *Cold Spring Harbor Symp. Quant. Biol.* **35**, 135–142.
Minkley, E. G. & Pribnow, D. (1973). *J. Mol. Biol.* **77**, 255–277.
Niles, E. G. & Condit, R. C. (1975). *J. Mol. Biol.* **98**, 57–67.
Oakley, J. L. & Coleman, J. E. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 4266–4270.
Oakley, J. L., Strothkamp, R. E., Sarris, A. H. & Coleman, J. E. (1979). *Biochemistry*. **18**, 528–537.
Osterman, H. L. & Coleman, J. E. (1981). *Biochemistry*. **20**, 4884–4892.
Pachl, C. A. & Young, E. T. (1978). *J. Mol. Biol.* **122**, 69–101.
Panayotatos, N. & Wells, R. D. (1979). *Nature (London)*. **280**, 35–39.
Pao, C. C. & Speyer, J. F. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 3642–3646.
Peters, G. G. & Hayward, R. S. (1974). *Eur. J. Biochem.* **48**, 199–208.
Pribnow, D. (1975). *J. Mol. Biol.* **99**, 419–443.
Robertson, H. D., Dixon, E. & Dunn, J. J. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 822–826.
Romano, L. J., Tamanoi, F. & Richardson, C. C. (1981). *Proc. Nat. Acad. Sci., U.S.A.*, **78**, 4107–4111.
Rosa, M. D. (1979). *Cell*. **16**, 815–825.
Rosa, M. D. (1981*a*). *J. Mol. Biol.* **147**, 55–71.
Rosa, M. D. (1981*b*). *J. Mol. Biol.* **147**, 199–204.
Rosa, M. D. & Andrews, N. C. (1981). *J. Mol. Biol.* **147**, 41–53.
Rosenberg, A. H., Simon, M. N., Studier, F. W. & Roberts, R. J. (1979). *J. Mol. Biol.* **135**, 907–915.
Rosenberg, M. & Court, D. (1979). *Annu. Rev. Genet.* **13**, 319–353.
Rosenberg, M. & Kramer, R. A. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 984–988.
Rosenberg, M., Kramer, R. A. & Steitz, J. A. (1974). *J. Mol. Biol.* **89**, 777–782.
Roth, J. R. (1981). *Cell*. **24**, 601–602.
Saito, H. & Richardson, C. C. (1981). *Cell*. **27**, 533–542.
Saito, H., Tabor, S., Tamanoi, F. & Richardson, C. C. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3917–3921.
Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A. III, Slocombe, P. M. & Smith, M. (1978). *J. Mol. Biol.* **125**, 225–246.
Scherzinger, E. & Seiffert, D. (1975). *Mol. Gen. Genet.* **141**, 213–232.
Scherzinger, E., Herrlich, P. & Schweiger, M. (1972). *Mol. Gen. Genet.* **118**, 67–77.
Shine, J. & Dalgarno, L. (1974). *Proc. Nat. Acad. Sci., U.S.A.* **71**, 1342–1346.
Siebenlist, U. (1979). *Nucl. Acids Res.* **6**, 1895–1907.
Siebenlist, U., Simpson, R. B. & Gilbert, W. (1980). *Cell*. **20**, 269–281.
Silberstein, S., Inouye, M. & Studier, F. W. (1975). *J. Mol. Biol.* **96**, 1–11.
Simon, M. N. & Studier, F. W. (1973). *J. Mol. Biol.* **79**, 249–265.
Stahl, S. J. & Chamberlin, M. J. (1977). *J. Mol. Biol.* **112**, 577–601.
Stahl, S. J. & Zinn, K. (1981). *J. Mol. Biol.* **148**, 481–485.
Steitz, J. A. (1980). In *Ribosomes Structure, Function and Genetics* (G. Chamblis *et al.*, eds), pp. 479–495. University Park Press, Baltimore.
Strome, S. & Young, E. T. (1978). *J. Mol. Biol.* **125**, 75–93.
Strome, S. & Young, E. T. (1980*a*). *J. Mol. Biol.* **136**, 417–432.
Strome, S. & Young, E. T. (1980*b*). *J. Mol. Biol.* **136**, 433–450.
Studier, F. W. (1969). *Virology*. **39**, 562–574.
Studier, F. W. (1972). *Science*. **176**, 367–376.

Studier, F. W. (1973*a*). *J. Mol. Biol.* **79**, 227–236.
Studier, F. W. (1973*b*). *J. Mol. Biol.* **79**, 237–248.
Studier, F. W. (1975*a*). *J. Mol. Biol.* **94**, 283–295.
Studier, F. W. (1975*b*). *Proc. 10th FEBS Meet.* 45–53.
Studier, F. W. (1975*c*). *J. Bacteriol.* **124**, 307–316.
Studier, F. W. (1979). *Virology*. **95**, 70–84.
Studier, F. W. (1981). *J. Mol. Biol.* **153**, 493–502.
Studier, F. W. & Rosenberg, A. H. (1981). *J. Mol. Biol.* **153**, 503–525.
Studier, F. W., Rosenberg, A. H., Simon, M. N. & Dunn, J. J. (1979). *J. Mol. Biol.* **135**, 917–937.
Summers, W. C. (1969). *Virology*. **39**, 175–182.
Summers, W. C. (1970). *J. Mol. Biol.* **51**, 671–678.
Sutcliffe, J. G., Shinnick, T. M., Green, N., Liu, F.-T., Niman, H. L. & Lerner, R. A. (1980). *Nature (London)*. **287**, 801–805.
Tamanoi, F., Saito, H. & Richardson, C. C. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2656–2660.
Walter, G., Scheidtmann, K.-H., Carbone, A., Laudano, A. P. & Doolittle, R. F. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 2311–2315.
Zaychikov, E. F. & Pletnev, A. G. (1980). *Bioorg. Khim.* **6**, 1268–1271.

Edited by M. Gottesman

Note added in proof: We have now found that purified T7 RNA polymerase initiates RNA chains with ATP at both the ϕOL and $\phi 2{\cdot}5$ promoters, as expected if RNA chains begin at position +1 of the conserved promoter sequence (Table 7 of the text). However, a minor but significant fraction of the chains were found to begin with GTP, presumably at position +2 of the promoter sequence. At ϕOL, perhaps 20% of the chains began with GTP, but at $\phi 2{\cdot}5$ the fraction was much smaller. In contrast, all of the RNA chains initiated at class III promoters have been found to begin with GTP at position +1 (Rosa, M. D. (1979). *Cell*. **16**, 815–825; Oakley, J. L., Strothkamp, R. E., Sarris, A. H & Coleman, J. E. (1979). *Biochemistry*. **18**, 528–537). T7 RNA polymerase clearly has a strong preference for initiating RNA chains at position +1 of the conserved promoter sequence, but changes from the class III promoter sequence apparently can allow some chain initiation at other positions.